Vocabulary Evolution on the Semantic Web

From Changes to Evolution of Vocabularies

and its Impact on the Data

Dissertation

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften

(Dr.-Ing.)

der Technischen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Mohammad Abdel-Qader

Kiel, 2020

Betreuer: Prof. Dr. Ansgar Scherp

Zweitgutachter: Prof.Dr. Wilhelm Hasselbring

Datum der Disputation: 30.09.2020

Zum Druck genehmigt: 03.02.2021

To my father, mother, wife, sons, brothers, and sisters
You are my rock, my biggest supporters, my everything.

Erklärung

Hiermit versichere ich,

1. dass diese Arbeit - abgesehen von der Beratung durch den Betreuer Prof. Dr. Ansgar Scherp - nach Inhalt und Form meine eigene ist,
2. dass Vorversionen einiger Teile dieser Arbeit bereits veröffentlicht wurden, nämlich

    - Mohammad Abdel-Qader and Ansgar Scherp. Qualitative analysis of vocabulary evolution on the linked open data cloud. In Elena Demidova, Stefan Dietze, Julian Szymanski, and John G. Breslin, editors, *Proceedings of the 3rd International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES '16) co-located with the 13th ESWC 2016 Conference, Anissaras, Greece, May 30, 2016*, volume 1597 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016
    - Mohammad Abdel-Qader, Ansgar Scherp, and Iacopo Vagliano. Analyzing the evolution of vocabulary terms and their impact on the LOD cloud. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 1–16. Springer, 2018
    - Mohammad Abdel-Qader, Iacopo Vagliano, and Ansgar Scherp. Analyzing the evolution of linked vocabularies. In Maxim Bakaev, Flavius Frasincar, and In-Young Ko, editors, *Web Engineering - 19th International Conference, ICWE 2019, Daejeon, South Korea,*

June 11-14, 2019, Proceedings, volume 11496 of *Lecture Notes in Computer Science*, pages 409–424. Springer, 2019

3. dass kein Teil dieser Arbeit bereits einer anderen Stelle im Rahmen eines Prüfungsverfahrens vorgelegen hat,
4. und dass diese Arbeit unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft entstanden ist.

_____

Mohammad Abdel-Qader

Acknowledgment

Foremost, I would like to express my sincere thanks to my supervisor, Prof. Dr. Ansgar Scherp. During my study, he was the teacher, the supporter, the guide, and the friend. I want to thank him for his patience, support for me at all stages of the study. I thank him for his tremendous knowledge. His guidance helped me in research and writing the thesis.

I have great pleasure in acknowledging my gratitude to my colleagues at Knowledge Discovery Group for their continued support and the wonderful times we spent together. Special thanks to Dr. Mahmoud Ghabashneh. Furthermore, I am very grateful to ZBW- The Leibniz Information Centre for Economics for the facilities they offered.

I would like to express my special thanks to all my friends. It would be inappropriate if I did not mention the names of Safwan Azab, Hatim Kittani, Abdalah Khodari, my brothers-in-law, and Hamzeh AbuTabanjah. They were my supporters during my studies, each in his own way, kept me going on my path to success and assisting me as per their abilities. Also, I would like my wife's family for their support.

I sincerely thanks my father Fakhri, my mother Nada, my lovely wife Muna, my sons Omar and Eyas, my brothers Saleh and Yazeed, my sisters Alaa and Razan. Thank you for supporting me spiritually throughout the study, writing this thesis and my life in general. Without your sincere support, I would not have been where I am today and what I am today.

Zusammenfassung

Das Hauptziel des Semantic Web ist es, den Daten im Web eine klar definierte Bedeutung zu geben. Vokabulare werden zum Modellieren von Daten im Web verwendet. Vokabulare vermitteln ein gemeinsames Verständnis einer Domäne und bestehen aus einer Sammlung von Typen und Eigenschaften. Diese Typen und Eigenschaften sind sogenannte Begriffe. Ein Vokabular kann Begriffe aus anderen Vokabularen importieren, und Datenverleger verwenden die Begriffe der Vokabulare zum Modellieren von Daten. Durch das Importieren von Begriffen entsteht ein Netzwerk verknüpfter Vokabulare (NeLO). Vokabulare können sich im Laufe der Zeit ändern. Wenn sich Vokabulare ändern, kann dies zu Problemen mit bereits veröffentlichten Daten führen, falls diese nicht entsprechend angepasst werden. Bisher gibt es keine Studie, die die Veränderung der Vokabulare im Laufe der Zeit analysiert. Darüber hinaus ist nicht bekannt, inwiefern bereits veröffentlichte Daten an diese Veränderungen angepasst werden. Verantwortliche für Ontologien und Daten sind sich möglicherweise der Änderungen in den Vokabularen nicht bewusst, da solche Änderungen eher selten vorkommen.

Diese Arbeit befasst sich mit dem Problem der Änderung von Vokabularen und deren Auswirkung auf andere Vokabulare sowie die Daten. Wir analysieren die Änderung von Vokabularen und deren Wiederverwendung. Für unsere Analyse haben wir diejenigen Vokabulare ausgewählt, die am häufigsten verwendet werden. Zusätzlich analysieren wir die Änderungen von 994 Vokabularen aus dem Verzeichnis "Linked Open Vocabulary". Wir analysieren die Vokabulare, um zu verstehen, von wem und wie sie in den modellierten Daten verwendet werden und inwiefern Änderungen in die Linked Open Data Cloud übernommen werden. Wir beobachten den Status von NeLO aus den verfügbaren Versionen der Vokabulare über einen Zeitraum von 17 Jahren. Wir analysieren statische Parameter von NeLO wie Größe, Dichte, Durchschnittsgrad und die wichtigsten Vokabulare zu bestimmten Zeitpunkten. Wir untersuchen weiter, wie sich NeLO mit der Zeit ändert. Insbesondere messen wir die Auswirkung einer Änderung in einem Vokabular auf andere, wie sich die Wiederverwendung von Begriffen ändert und wie wichtig Änderungen im Vokabular sind.

Unsere Ergebnisse zeigen, dass die Vokabulare sehr statisch sind und viele Änderungen an sogenannten Annotations-Properties vorgenommen wurden. Darüber hinaus werden 16% der vorhandenen Begriffen von anderen Vokabularen wiederverwendet, und einige der veralteten und gelöschten Begriffe werden weiterhin wiederverwendet. Darüber hinaus werden die meisten neu erstellten Begriffe unmittelbar verwendet. Unsere Ergebnisse zeigen, dass selbst wenn die Häufigkeit von Änderungen an Vokabularen eher gering ist, so kann dies aufgrund der großen Datenmenge im Web erhebliche Auswirkungen haben. Darüber hinaus hat sich aufgrund einer großen Anzahl von Vokabularen in NeLO und damit der Zunahme der verfügbaren Begriffe der Prozentsatz der importierten Begriffe im Vergleich zu den verfügbaren Begriffen im Laufe der Zeit verringert. Basierend auf den Ergebnissen der durchschnittlichen Anzahl von Exporten für die Vokabulare in NeLO sind einige Vokabulare im Laufe der Zeit immer beliebter geworden. Insgesamt ist es für Verantwortliche für Ontologien und Daten wichtig, die Entwicklung der Vokabulare zu verstehen, um falsche Annahmen über die im Web veröffentlichten Daten zu vermeiden. Darüber hinaus ermöglichen unsere Ergebnisse ein besseres Verständnis der Auswirkungen von Änderungen in Vokabularen, sowie deren Nachnutzung, um möglicherweise aus früheren Erfahrungen zu lernen. Unsere Ergebnisse bieten erstmals detaillierte Einblicke in die Struktur und Entwicklung des Netzwerks der verknüpften Vokabularen. Unterstützt von geeigneten Tools für die Analyse in dieser Arbeit, kann es Verantwortlichen für Ontologien helfen, Mängel in der Datenmodellierung zu identifizieren und Abhängigkeiten zu bewerten, die durch die Wiederverwendung eines bestimmten Vokabulars entstehenden.

Abstract

The main objective of the Semantic Web is to provide data on the web well-defined meaning. Vocabularies are used for modeling data in the web, provide a shared understanding of a domain and consist of a collection of types and properties. These types and properties are so-called terms. A vocabulary can import terms from other vocabularies, and data publishers use vocabulary terms for modeling data. Importing terms via vocabularies results in a Network of Linked vOcabularies (NeLO). Vocabularies are subject to change during their lifetime. When vocabularies change, the published data become a problem if they are not updated based on these changes. So far, there has been no study that analyzes vocabulary changes over time. Furthermore, it is unknown how data publishers reflect on such vocabulary changes. Ontology engineers and data publishers may not be aware of the changes in the vocabulary terms that have already happened since they occur rather rarely.

This work addresses the problem of vocabulary changes and their impact on other vocabularies and the published data. We analyzed the changes of vocabularies and their reuse. We selected the most dominant vocabularies, based on their use by data publishers. Additionally, we analyzed the changes of 994 vocabularies. Furthermore, we analyzed various vocabularies to better understand by whom and how they are used in the modeled data, and how these changes are adopted in the Linked Open Data cloud. We computed the state of the NeLO from the available versions of vocabularies for over 17 years. We analyzed the static parameters of the NeLO such as its size, density, average degree, and the most important vocabularies at certain points in time. We further investigated how NeLO changes over time, specifically measuring the impact of a change in one vocabulary on others, how the reuse of terms changes, and the importance of vocabulary changes.

Our results show that the vocabularies are highly static, and many of the changes occurred in annotation properties. Additionally, 16% of the existing terms are reused by other vocabularies, and some of the deprecated and deleted terms

are still reused. Furthermore, most of the newly coined terms are adopted immediately. Our results show that even if the change frequency of terms is rather low, it can have a high impact on the data due to a large amount of data on the web. Moreover, due to a large number of vocabularies in the NeLO, and therefore the increase of available terms, the percentage of imported terms compared with the available ones has decreased over time. Additionally, based on the scores of the average number of exports for the vocabularies in the NeLO, some vocabularies have become more popular over time. Overall, understanding the evolution of vocabulary terms is important for ontology engineers and data publishers to avoid wrong assumptions about the data published on the web. Furthermore, it may foster a better understanding of the impact of the changes in vocabularies and how they are adopted to possibly learn from previous experience. Our results provide for the first time in-depth insights into the structure and evolution of the NeLO. Supported by proper tools exploiting the analysis of this thesis, it may help ontology engineers to identify data modeling shortcomings and assess the dependencies implied by the reusing of a specific vocabulary.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The main objective of the Semantic Web is to provide data on the web well-defined meaning. This allows computers apart from humans to understand the data. These meanings can be represented using vocabularies. Gruber [25] defined vocabularies as a descriptive form for concepts and items in a specific field or domain, providing specifications for those items and their relations to other concepts. Vocabularies consist of a collection of types and properties known as terms and published as Linked Data. Linked Data is structured data that has links to other datasets in order to discover more information through semantic tools. In 2001, Tim Berners-Lee introduced the idea of the Semantic Web, and in 2006 summarized the Linked Data principles into four rules [9].

1. Name things (resources) using a Uniform Resource Identifier (URI).

2. Allow people to look for those names by using HTTP URIs as unique names for resources.

3. Use the Semantic Web standards to provide useful information when someone looks up the URI.

4. Make a connection to other URIs. Thus, more information can be detected.

Publishing data on the web forms the Linked Open Data (LOD) cloud. The LOD cloud is a graph that represents connected datasets that publish data in the form of Linked Data. As of March 2019, the LOD cloud consists of 1,239 datasets and 16,147 links between those datasets[1]. Figure 1.1 shows the LOD cloud as of March 2019.

Vocabularies are used to model data based on the principles of Linked Data. The Semantic Web standards such as the Resource Description Framework (RDF) [46], the RDF Vocabulary Definition Language (*rdfs*) [11], and the Web Ontology Language (*owl*) [51] are considered the basis for establishing the other vocabularies. Vocabularies can be connected when ontology engineers reuse terms from other vocabularies, and these connections can define mappings to other related vocabularies [9].

This thesis focuses on analyzing the evolution of vocabularies on the Semantic Web. After ontology engineers published a version of vocabulary, the terms are subject to changes [44]. We aim to provide a better understanding of the changes of vocabularies and the relation between them and their importers, i. e., the vocabularies that reuse one or more terms from other vocabularies. Furthermore, we attempt to assist data publishers in updating their models by providing information about the current status of terms, statistics regarding the use of vocabulary types and properties, and provide analyses about vocabulary

[1]The Linked Open Data cloud diagram 2019 is made by John P. McCrae, Andrejs Abele, Paul Buitelaar, Richard Cyganiak, Anja Jentzsch, Vladimir Andryushechkin, and Jeremy Debattista. `http://lod-cloud.net/`, last accessed November 28, 2019

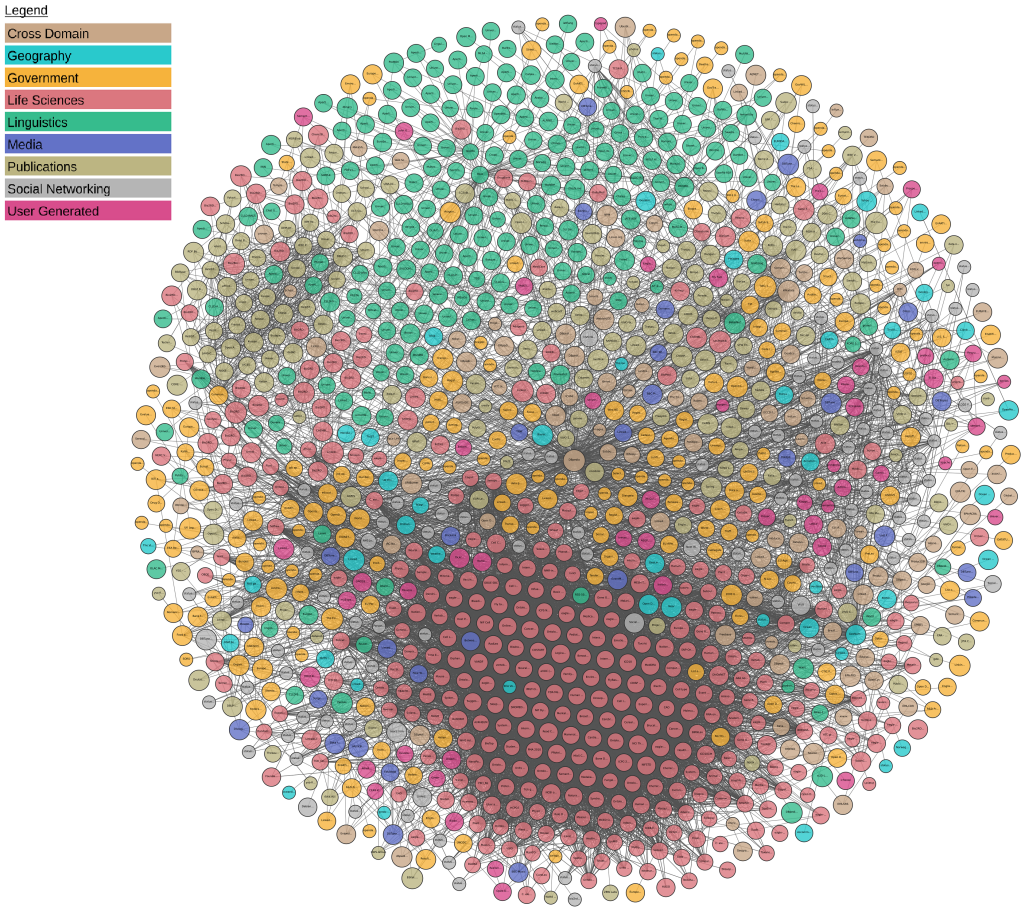


Figure 1.1: The Linked Open Data (LOD) cloud diagram as of March 2019. The circles are the datasets, and their colors specify the domain to which they are related.

changes. Additionally, we intend to provide a better understanding of the impact of vocabulary changes on data and other vocabularies, as well as aim to analyze how data publishers adopt the changed terms.

We analyze the changes in vocabularies and the reuse of vocabulary terms by others. We extract the imported types and properties from other vocabularies to check if they have been deleted or deprecated. We analyze all the available versions of the vocabulary and record the terms that were deprecated or deleted and check whether other vocabularies still import them. Subsequently, we analyze the use of vocabulary terms in the published data as well as the adoption of vocabulary changes by data publishers. We also analyze various vocabularies to better understand by whom and how they are used, and how data publishers adopt these changes. Finally, we analyze the evolution of the NeLO, by employing a broad range of network-analysis metrics on the generated network and then applying them during the evolution to find out how the important nodes have changed over time.

The results show that most of the vocabularies are highly static, in line with a study by Käfer et al. [37]. Additionally, the results show that the ontology engineers may not be aware of changes in the vocabularies exploited, and most of the newly coined terms are adopted in the published data in less than a week. Moreover, after analyzing the NeLO, we found that the amount of reuse is low. Therefore, there is a need to increase the import/export relations between vocabularies. Finally, based on the average number of exports of the vocabularies in the NeLO, some vocabularies have become more popular over time.

The remainder of the chapter is structured as follows. In Section 1.1, we introduce the motivating examples for the work. Subsequently, the research questions

and their contributions are outlined in Section 1.2. In Section 1.3, we list the publications before we outline the remainder of this thesis.

1.1 Motivation

One of the key principles for modeling and publishing data on the web is to reuse terms, i. e., types and properties from existing vocabularies. The types or classes are used to describe the type of resources, and the properties represent the relations between them. Vocabulary terms are reused by other vocabularies and data publishers to model their data. We consider the term to be "reused" when it is imported by other vocabularies. The "use" of term is when the data publishers employ the considered term to model data. The terms are subject to change. Therefore, the reused terms by other vocabularies, and the used ones in the published data, must be revised to check for reused changed terms, and if they have been changed. The connections between vocabularies and the published data of the LOD cloud motivate us to analyze the changes of vocabularies and the impact of these changes on other vocabularies and the published data. Below are three motivating examples that describe the scenarios regarding the following:

1. The reuse of vocabulary terms by other vocabularies and the impact of vocabulary changes.

2. The use of vocabulary terms in the published data and the adoption of changes in vocabularies by data publishers.

3. The evolution of the NeLO

Reuse of Vocabulary Terms and the Impact of their Changes

Regarding the impact of vocabulary changes on other vocabularies that reuse one or more of its types and properties, we consider the example of the Asset Description Metadata Schema (*adms*) vocabulary. It deals with describing highly reusable metadata and reference data known as Semantic Assets[2]. Figure 1.2 (bottom) shows the evolution of the *adms* vocabulary over six versions within five years, between May 2012 and July 2015. Figure 1.2 (top) depicts the *food* vocabulary, which reuses some terms of *adms* and also had different versions over its lifespan. The *adms* vocabulary was introduced as the `adms:SemanticAsset` type and `adms:accessURL` property in its version published in June 2012 (V2). The *food* vocabulary imported those two terms in its first version, which was published in November 2012. Subsequently, a new version of the *adms* vocabulary was released in May 2013, which deleted the `adms:SemanticAsset` and `adms:accessURL` terms. In September 2013, the *food* vocabulary was updated, but the updated version of the *food* vocabulary keeps reusing the two terms that were deleted from the *adms*. Such a scenario may mean that the *food* vocabulary still needs the deleted terms and its ontology engineers have found no alternatives. However, it could also indicate that the ontology engineers of the *food* vocabulary are not aware of the changes in the *adms*.

[2] `https://www.w3.org/TR/vocab-adms/`

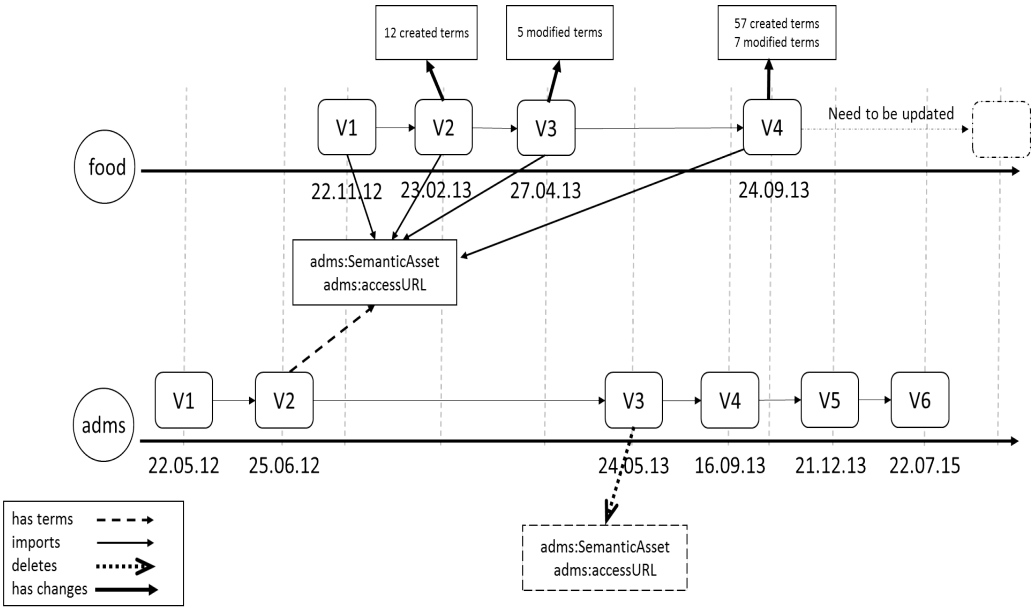


Figure 1.2: The evolution of the *adms* vocabulary and the impact of change on one of its importers, the *food* vocabulary.

Use and Adoption of Vocabulary Terms for Modeling Data

Vocabularies are used for modeling data in the LOD cloud and Wikidata. During their lifetime, vocabularies are subject to changes; new terms are coined, and existing terms are modified or deprecated. Data publishers may not be aware of changes in the vocabulary terms since they occurs rather rarely [37]. Explicitly triggering data publishers to update their model is also challenging due to the distributed nature of published data. However, in general, data providers may be interested in being notified when specific vocabulary term changes happen. Until recently, data providers lacked the proper tools and services to track whether and the kind of changes in vocabulary terms have occurred.

The "use" of a vocabulary term means that the term has appeared in the published LOD cloud, while the "adoption" of terms is the use of newly created terms in the published data. Figure 1.3 shows an example of the *adms* vocabulary

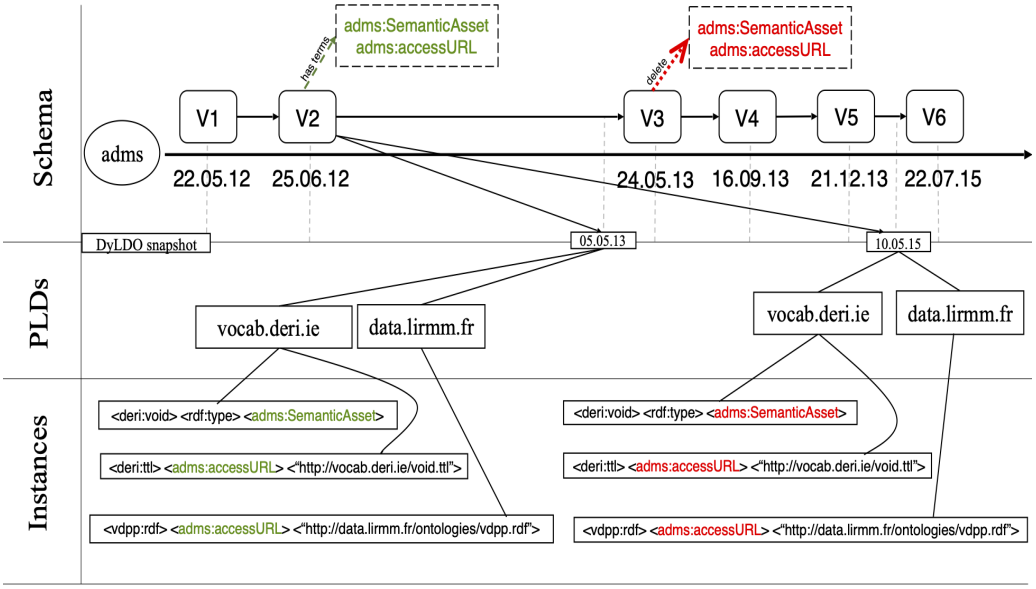


Figure 1.3: The evolution of the *adms* vocabulary and the impact of its change on the published data.

evolution and the use of some of its terms in the published data on the LOD cloud. In May 2012, the *adms* vocabulary introduced the `adms:SemanticAsset` type and the `adms:accessURL` property. In early May 2013, the previous two terms appeared in a snapshot of one of the LOD cloud datasets, namely, the Dynamic Linked Data Observatory [38] dataset. Thus, the newly created terms got adapted. The two terms were used by two Pay Level Domains (PLDs). PLD refers to any web domain that requires payment at a Top Level Domain (TLD) or country code TLD (cc-TLD) registrar [47]. The two PLDs are `vocab.drei.ie` and `data.lirmm.fr`. In late May 2013, the *adms* got a new version release where `adms:SemanticAsset` and `adms:accessURL` were deleted. The PLDs `vocab.drei.ie` and `data.lirmm.fr` PLDs still use the two terms that were deleted from *adms*, two years after their deletion date. This example shows that the published data may need to be updated to reflect the changes in vocabularies.

The Network of Linked Vocabularies

Figure 1.4 shows a selected part of the NeLO. This network is formed when vocabularies reuse terms from other vocabularies which creates connections to other vocabularies. Thus, the nodes in this network are the vocabularies. Furthermore, some of the dependencies of the vocabularies are also depicted in this figure. The arrows represent the relations between the vocabularies. An arrow from a vocabulary W to another vocabulary V indicates that V imports terms from W, or, in other words, that W exports terms to W. The size of the nodes represents the number of exports, i. e., more exports imply a bigger node. Additionally, the width of the edges represents the total number of types and properties that are imported by the target vocabulary from the source vocabulary. For example, the *adms* vocabulary exports types and properties to the following vocabularies: *food*, *gn*, *search*, and *void*. On the other hand, *schema*, and *voaf* export terms to the *adms*.

From the example in Figure 1.4, we can see that there are edges between vocabularies when some reuse types or properties from others. Therefore, a change in one vocabulary of the network will affect the other vocabularies that import terms from the changed vocabulary.

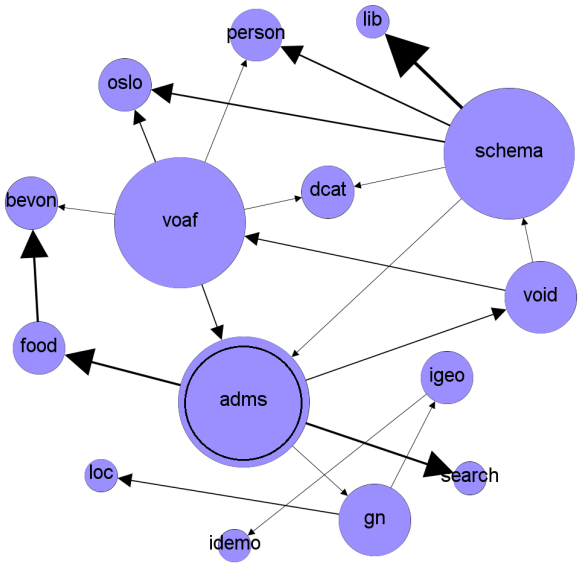


Figure 1.4: Selected excerpt of the Network of Linked Vocabularies (NeLO) showing vocabularies that import types or properties from the *adms* vocabulary and other vocabularies.

1.2 Research Questions

The three scenarios introduced in Section 1.1 motivated us to explore and analyze the evolution of vocabularies on the Semantic Web. From them, we derived the main research question of this thesis:

> *How do the changes in vocabularies affect the data on the Linked Open Data cloud and the other vocabularies of the Network of Linked Vocabularies?*

We divided the main research question into three research questions, RQ1 to RQ3, which makes achieving our main goal possible by solving the sub-problems separately. For RQ1, we analyzed the changes in vocabularies and the reuse of terms between vocabularies. For RQ2, we analyzed the use of terms in the LOD cloud and the adoption of vocabulary changes by data publishers. For RQ3, we

analyzed the evolution of the NeLO and the impact of vocabulary changes on the vocabularies of the network.

Research Question 1 (RQ1): Vocabulary Changes and the Reuse of Terms by other Vocabularies

Motivation. Vocabulary terms are reused by data publishers to model their data. One of the key principles for modeling and publishing data on the web is to reuse types and properties from existing vocabularies.

Problem Statement. Vocabularies are subject to change. The changes are needed to reflect the changes into vocabulary's domain, to resolve errors in prior versions, or to reflect the changes in terms of the imported vocabularies. For this research question, we focused on analyzing the changes that have occurred on vocabularies. We analyzed the vocabularies from different perspectives to observe how changes and types of it in one vocabulary are influencing the vocabularies that reuse terms from the changed vocabulary. Possible changes are addition, deletion, and deprecation of terms. Any other change, e. g., a modification, can be expressed using these two basic changes. We also analyzed the vocabularies to study their reuse and some of the deleted terms imported from other vocabularies. Furthermore, we analyzed the number of terms reused by other vocabularies. RQ1 includes several subquestions:

- How are vocabularies influenced by changes made in their dependent vocabularies?
- How do vocabularies reuse terms imported from other vocabularies?

- How many vocabularies import terms from the other vocabularies?
- How many terms are imported by vocabularies? Are the imported terms the most recent ones?

Procedure. To collect the required information about vocabulary changes, we used the Linked Open Vocabularies (LOV)[3] service to download all the available versions of the vocabularies. LOV is an online platform to represent the dependencies of linked open vocabularies. This allowed us to look back at some versions of the vocabularies to more than 17 years. We considered vocabularies as part of the network if they imported or exported at least one type or property from some other vocabulary. We extract the imported types and properties from other vocabularies to check if they had not deleted or deprecated. We parsed all the versions of the vocabularies and record the terms that were deprecated or deleted. Subsequently, we check if they are still imported by other vocabularies.

Contribution to RQ1. To study the changes of vocabularies and the reuse of vocabulary terms by others, we analyzed a large set of vocabularies to provide a comprehensive analysis. The analysis shows that some of the deprecated/deleted types and properties are still reused by other vocabularies. We found that 33 of the vocabularies in the LOD cloud, as of June 2018 still reuse deprecated or deleted terms. The analysis shows that in the early versions of vocabularies, most of the changes occurred on annotation properties. This reflects a need for more clarification of the terms. Additionally, the analysis also shows that not all vocabulary terms are reused by other vocabularies, and present the most reused types

[3] `http://lov.okfn.org/dataset/lov/`, last accessed: November 28, 2019

and properties. Only 16% of the existing terms are reused by some other vocabularies, based on the vocabularies listed in LOV until June 2018.

Research Question 2 (RQ2): Use and Adoption of Vocabulary Terms for Modeling Data

Motivation. Vocabulary terms define the schema of the LOD cloud or Wikidata. The LOD cloud is a graph that represents connected datasets that publish data in the form of Linked Data. Wikidata is a knowledge base to collaboratively store and edit structured data; it is a free and multilingual repository[4]. After ontology engineers published the first version of a vocabulary, the terms are subject to changes to reflect new requirements or shifts in the domains of the vocabularies models. Thus, data modeled on these vocabularies need to be updated, too.

Problem Statement. So far, it is unknown how the data publishers adopt the newly coined terms of vocabularies to model their data. They may not be aware of the changes in the vocabulary terms, since they occurs rather rarely [37]. Explicitly triggering data publishers to update their model is also challenging due to the distributed nature of the LOD cloud. Data publishers may be interested in being notified when specific vocabulary term changes happen, but they lack the proper tools and services to track whether and what kind of changes have occurred in vocabulary types and properties. Likewise, ontology engineers are not aware of who uses their vocabularies and lack tools that reflects the adoption status of their ontologies and changes on the defined terms. "Adoption" is the first use of the newly created terms. The difference between the use and adoption

[4] `https://www.wikidata.org/,lastaccessed:November28,2019`

of vocabulary terms is that vocabulary term use refers to the analysis of the use of terms over time in the published data of the LOD cloud, while vocabulary term adoption refers to the use of newly created terms in the published data. Furthermore, data publishers that use types and properties of a changed vocabulary should update their data according to those changes. This is especially needed when a term is deleted. For this research question, we studied the use of vocabulary terms for modeling data, particularly when data publishers adopted the newly created terms. Accordingly, we address three research subquestions:

- How many times are terms from a specific vocabulary used by a dataset?
- Are the deleted and deprecated types and properties still used by data publishers?
- When are the newly created types and properties adopted by data publishers, i. e., first used?

Procedure. We studied the relationship between vocabulary changes and the published data, and analyzed various vocabularies to better understand by whom and how they are used, and how these changes are adopted in the evolving LOD cloud. We considered the three basic types of changes: addition, deletion, and deprecation. Any other change, e. g., a modification, can be expressed through these three basic changes. We used three datasets of crawled data from the LOD cloud: DyLDO [38], BTC[5], and Wikidata[6].

The first dataset is a collection of weekly snapshots for a set of linked data documents from the LOD cloud. The second is yearly crawled from the

[5] `http://challenge.semanticweb.org/`, last accessed: November 28, 2019
[6] `https://www.wikidata.org`, last accessed: November 28, 2019

LOD cloud from 2009 to 2012, as well as in 2014. From both datasets, we extracted the Pay Level Domains (PLDs) that adopted the changed terms of vocabularies. We extracted the vocabulary types and properties from Wikidata and determined whether changes in the vocabulary were done (additions, deletions or deprecations), and how these changes were adopted. We observed the adoption of vocabulary types and properties after new versions of the vocabularies were published.

Contribution to RQ2. We studied the relationship between vocabulary changes and the published data. Our experiments show that even if the change frequency of vocabulary terms is rather low, they have a large impact on the published data. Most of the newly coined terms are adopted in less than one week after their publishing date. However, some types and properties are only adopted after several months or a few years after the date of creation, while some other adoptions happen even before the official publishing data of a term. Many deprecated terms are still in use in the published data. Our results show that for most vocabularies, notably in the BTC dataset, more than 50 % of types and properties are actually unused. We provide an analysis that can foster a better understanding of the impact of the vocabulary changes and how terms are adopted. For example, we believe the analysis can make ontology engineers more aware of who uses their terms and how. This analysis can also assist data publishers in updating their models by providing information about vocabulary changes.

Research Question 3 (RQ3): Analysis of the Network of Linked Vocabularies

Motivation. It is common practice to reuse existing terms, i. e., properties and types, defined in the vocabularies to build other vocabularies. This reuse of terms leads to a NeLO. In essence, NeLO is a directed graph of connected vocabularies that contain at least one reuse from some other vocabulary. By connected vocabularies, we mean that a vocabulary V is importing at least one type or property from another vocabulary W. For example, in Section 1.1, we showed an example that depicts a part of the NeLO, where the Sindice Search Vocabulary (*search*), Geonames (*gn*), the Food Ontology (*food*), and the Vocabulary of Interlinked Datasets (*void*) import some terms from the *adms* (Figure 1.4).

Problem Statement. The connections between the vocabularies become a problem when one or more of the vocabularies in the network change. For example, when the vocabulary W declares a term t as deprecated or even deletes it, but the dependent vocabulary V is importing this term t. The changes of vocabularies have a direct impact on all dependent vocabularies, i. e., those that import any of the changed terms. Furthermore, all the data that are modeled on these outdated vocabularies have to be updated, too.

To the best of our knowledge, there is no study about the evolution of the NeLO, i. e., how it changes over time. Previous researches have focused on analyzing the interlink at an instance level, i. e., the interlink between the published data. For example, the LOD cloud[7] (Figure 1.1) studies the interconnection among datasets. In contrast to RQ2, for analyzing the NeLO, we focused on the evolution of the web of data at the schema level.

[7] `http://lod-cloud.net/`

We analyzed the changes in the NeLO by addressing the following research subquestions:

- What is the state of the NeLO as of June 2018?
  - What is the size concerning the number of nodes (vocabularies) and edges (relations)?
  - What is its density, and average degree?
  - What are the crucial vocabularies, i. e., central nodes?
- How does the change of terms in one vocabulary impact the other vocabularies on the network?
- How do ranking metrics, such as *PageRank*, *Hypertext Induced Topic Selection* (HITS), and *Centrality*, change during the evolution of the NeLO?

Procedure. We analyzed the evolution of the NeLO over time, starting from its early stages, when just a few vocabularies appeared (2001), to the time the number of vocabularies almost doubled (2008), until the big growth of the network (2018). Using the LOV dataset, we downloaded all the available vocabularies and their versions until June 2018, and subsequently extracted all types and properties from all versions of the vocabularies. We employed a broad range of network-analysis metrics on the generated network and applied them to the evolution of the NeLO to find out how the important nodes change over time. Additionally, we investigated how the change of one vocabulary impacts the others that import its terms.

Contribution to RQ3. To analyze the evolution of the NeLO, we used the

LOV dataset. We downloaded all the available vocabularies and their versions until June 2018, providing a comprehensive analysis. From the figure of NeLO as of June 2018, we concluded that the vocabularies are organized into three categories. The inner circle mostly consists of the meta-vocabularies. The middle circle includes the highly popular vocabularies but not like the meta-vocabularies. The outer circle is the one that contains the rarely used vocabularies and the newcomers. We believe that most of the newcomers in the future will be in the outer circle for a while, since the more general and the meta-vocabularies (the center circle) are already covering a broad range of data modeling needs, and we do not expect that there will be many new generic vocabularies in the future. Furthermore, our analysis shows that the NeLO was large at the beginning of the growth, but recently, the increase has been slower. While the growth rate was 43% in June 2010, it has decreased to only 4% as of June 2018 in terms of the number of nodes and edges. We provide, for the first time, in-depth insights into the structure and evolution of NeLO over 17 years, and a yearly graph from the year 2001 to 2018.

1.3 Publications List

This thesis is based on three publications, published in two conferences and one workshop. The three publications are as follows:

- Mohammad Abdel-Qader and Ansgar Scherp. Qualitative analysis of vocabulary evolution on the linked open data cloud. In Elena Demidova, Stefan Dietze, Julian Szymanski, and John G. Breslin, editors, *Proceedings of the 3rd International Workshop on Dataset PROFIling and fEderated*

Search for Linked Data (PROFILES '16) co-located with the 13th ESWC 2016 Conference, Anissaras, Greece, May 30, 2016, volume 1597 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016

- Mohammad Abdel-Qader, Ansgar Scherp, and Iacopo Vagliano. Analyzing the evolution of vocabulary terms and their impact on the LOD cloud. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 1–16. Springer, 2018

- Mohammad Abdel-Qader, Iacopo Vagliano, and Ansgar Scherp. Analyzing the evolution of linked vocabularies. In Maxim Bakaev, Flavius Frasincar, and In-Young Ko, editors, *Web Engineering - 19th International Conference, ICWE 2019, Daejeon, South Korea, June 11-14, 2019, Proceedings*, volume 11496 of *Lecture Notes in Computer Science*, pages 409–424. Springer, 2019

1.4 Thesis Outline

In Chapter 2, we review the related work. In Section 2.1 of the related work chapter, we start by providing a brief introduction of the Semantic Web principles and defining all the terminologies related to it, relevant in this thesis. Subsequently, we review the works that have analyzed the changes of vocabularies and the reuse of terms by other vocabularies of the Semantic Web (Section 2.2 reviews the studies that analyzed the changes of vocabularies and

Section 2.3 reviews the works that have studied the reuse of vocabulary terms), which are related to RQ1. Section 2.4 reviews the works related to RQ2, which have analyzed the impact of vocabulary changes on the published data in the LOD cloud. Subsequently, we reviewed the works related to the evolution of the NeLO (RQ3) and the analyses conducted in this area (Section 2.5).

Chapters 3, 4, and 5 represent the three research questions RQ1, RQ2, and RQ3, respectively. Chapter 3 describes the methodology conducted to extract the changes of vocabularies and the reuse of vocabulary types and properties by other vocabularies, the datasets used to conduct the experiments, the results found using that methodology, and the discussion of the results.

Subsequently, we describe the methodology for analyzing the reuse of vocabulary types and properties by data publishers and the adoption of vocabulary changes in the LOD cloud in Chapter 4. This chapter contains a description of the datasets used to analyze the reuse of vocabulary terms and the adoption of the newly created terms. Subsequently, we list the results before we discuss them.

The analyses of the changes in the NeLO are discussed in details in Chapter 5, which includes the procedure, analysis metrics, results of the evolution of the NeLO, the analysis of its state as of June 2018, and the discussion of the results.

Chapter 6 concludes the main contributions regarding the three research questions, lists the lessons learned from the challenges we faced during answering them, and provides future directions and outlines for further investigations.

Chapter 2

Related Work

In this chapter, we introduce the Semantic Web principles needed for this thesis and define the terminologies related to the Semantic Web (Section 2.1), which have been used in this study. In Sections 2.2 and 2.3, we review the related work that have analyzed the reuse of vocabulary terms between vocabularies and the changes of vocabularies. Section 2.4 reviews the works that have analyzed the use of vocabulary terms by data publishers and how vocabulary changes are adopted in the LOD cloud. A literature review for the works that have analyzed the evolution of the NeLO are introduced in Section 2.5, before we summarize.

2.1 Principles of the Semantic Web and its Terminologies

In this section, we briefly explain some of the principles of the Semantic Web. Furthermore, we define the terminologies used in this thesis. Please note that this

section will not cover all the Semantic Web principles, but only the ones that are needed for the readers of this work. For more details, there are several resources on the field of Semantic Web, such as [40, 43, 28].

Semantic Web: It is an extension of the current web. The information in the web is given a well-defined meaning (semantics) in order to allow machines beside humans to understand the meaning of information [8, 40].

Linked Open Data (LOD): It is the intersection between the idea of Linked and Open Data. Linked Data are structured data that have links to other datasets in order to discover more information through semantic tools. The main idea is to make a web of related data rather than a web of related documents [9]. Open Data refers to data that is freely available to everyone without restrictions [43]. Open Data may or may not be linked to other datasets and must be in a machine-readable format. The data is open to access and linked to other data [9].

LOD is five-star data as shown in Figure 2.1[1]. Tim Berners-Lee defined the LOD as[2]:

> *"Linked Data which is released under an open license, which does not impede its reuse for free"*

The data becomes one-star data when it is available on the web as open access data. The data can be in any format. When the data becomes structured data, in addition to the one-star data benefits, it obtains a two-star rating. A third star can be obtained if the data is two-star plus available in a non-proprietary (open) format, such as *csv* and *xml*. The three-star

[1] `https://5stardata.info/en/`, last accessed: November 28, 2019
[2] `https://www.w3.org/DesignIssues/LinkedData.html`

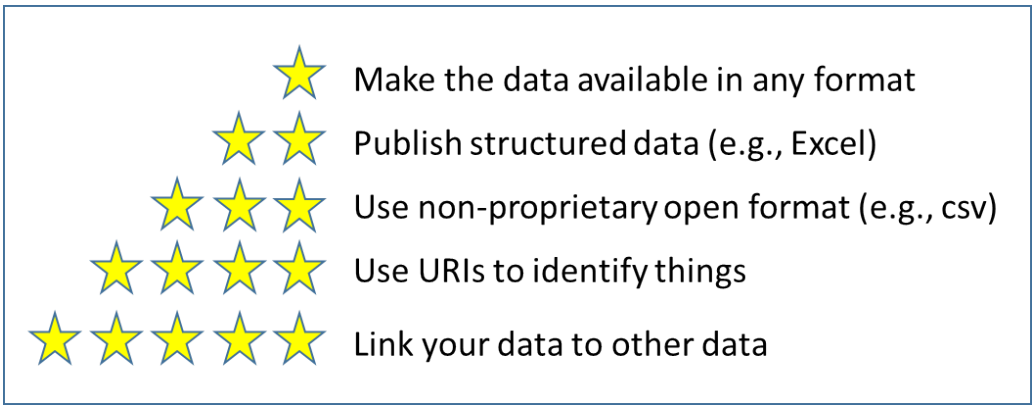


Figure 2.1: The 5-star open data plan.

data can be accessed with different software systems. When the resources (things) are given a unique ID, then the data becomes four-star. This unique ID is called as a Uniform Resource Identifier (URI). The five-star data is achieved when it is linked to other data to allow users to discover more information.

***RDF* Triple:** An *RDF* triple Tr is defined as a triple $Tr = (s, p, o)$, where s is the subject part, p is the predicate part, and o is the object part of a triple [28]. The triple is represented as:

<Subject> <Predicate> <Object>

The following is an example of a statement and how it is represented as a triple:

John has a car
<John> <has a> <car>

Subject: It can be either a URI or a blank node. A blank node represents a resource without a URI or literal [31]. Literal represents values such as

names, numbers, and dates[3]. The subject is the primary resource being described. In this example, the subject is *"John"*.

Predicate: It must be a URI, which represents the relation between the subject and the object parts of a triple. In the example, the predicate is *"has a"*.

Object: It can be a URI, blank node, or literal, which represents the value of the relation with the subject part. In the example, the object is *"car"*.

Since we focused on studying the changes in vocabulary terms, please note that we only considered the resources and predicates that are identified by a URI for this work.

Vocabulary (v): Formally, we understand a vocabulary v as a set of terms T; v is a set of well-defined terms (types and properties) that have a broad meaning and used to describe things (resources) [30]. These vocabularies can be general, which are suitable for all domains, or specific for some or a single domain[4].

Vocabulary term: A term t is either a class C or a property P. A set of terms relates to vocabulary v as $T(v) = C(v) \bigcup P(v)$. A class C is also known as type or concept. A class C is used to describe the semantic type of a resource. A property P is used to describe the semantic meaning of the relationship between resources or the resource and the literal.

Meta vocabularies: There is no precise classification of the meta-vocabularies,

[3] `https://www.w3.org/TR/rdf11-s/`, last accessed: November 28, 2019
[4] `https://www.w3.org/standards/semanticweb/ontology`, last accessed: November 28, 2019

but we can define them as the most general vocabularies that are used to construct most of the other vocabularies. In this thesis, we consider the*owl*, the *rdf*, the rdf Schema (*rdfs*), the eXtensible Markup Language (*xml*), and the xml Schema (*xsd*).

Outdated term: An outdated term t' is the type (class) C or property P that was deleted or deprecated when updating a vocabulary. The deleted term is the one that is excluded from the updated version. The deprecated term is the one that still exists in the updated version but is marked as outdated, i. e., it should not be used anymore.

Reused term: Formally, we consider that the term t of vocabulary W is reused by vocabulary V when it imports the term t of vocabulary W. Thus, this refers to a vocabulary reuse, but not necessarily that the term is used in any data.

Used term: The term t is considered to be used if it is employed for modeling some data, i. e., the vocabulary term t occurs in at least one triple Tr in a published dataset.

Adoption of term: The adoption of term t is the use of newly created terms in the published data.

Pay Level Domain (PLD): A PLD is any domain that requires payment at a Top Level Domain (TLD) or country coded TLD (cc-TLD) registrar. PLDs are usually one level below the corresponding TLD (e.g., amazon.com), with certain exceptions for cc-TLDs (e.g., ebay.co.uk, det.wa.edu.au) [47].

Network of Linked Vocabularies (NeLO): Formally, we see the NeLO as a

directed graph $G = (V, E)$, where V is the set of vertices, or nodes, and E is the set of edges, or links. In our case, the nodes are the vocabularies, and the directed edge $(v, v') \in E$ from a vocabulary v to another v' means that v' imports at least one term t from v, with both v and v' belonging to V.

2.2 Changes of Vocabularies on the Linked Open Data Cloud

In this section, we first review the works that have analyzed changes of vocabularies and how these changed vocabularies are used in the LOD cloud. Then, we present the studies that have discussed the process of ontology design and evolution from the perspective of ontology engineers. Finally, we review the works related to the analysis of vocabulary evolution.

The research so far has focused on analyzing the changes of the data, but not how the vocabularies change. For example, Dividino et al. [18] proposed a framework to measure the evolution of the data in a dataset over time. They proposed a dynamics function and applied it to 84 weekly snapshots from the DyLDO dataset, to compute a value representing how much the data in the dataset had evolved. The DyLDO dataset is a repository to store weekly snapshots from a subset of the LOD. Neubert [54] compared the versions of *skos* vocabulary and stored the differences in two named graphs representing the insertions and deletions. Then, he stored metadata that describes versions and deltas in a separate file. He called this file a version history graph.

Walk et al. [73] studied the user behavior during the process of editing ontologies in order to improve the support of ontology engineering tools. They derived

nine hypotheses to describe the change behavior of users, and then applied those hypotheses on four real-world ontology projects. They found that the hierarchical structure hypothesis had the most substantial influence on the editing behavior. It showed that the users edit classes along with hierarchical relations (parent, child, sibling, cousin). Walk et al. [74] analyzed the collection of actions in the change-logs made by users in the collaborative ontology engineering methods and techniques to increase the quality of ontologies. They applied Markov chains to the International Classification of Diseases (Revision 11) dataset. The authors provided description of the process for applying Markov chains on the change-logs, as well as an evaluation of the extracted Markov chain models. Zablith et al. [75] published a survey presenting an ontology evolution cycle, trying to gather researchers' work in the ontology evolution community. Furthermore, they analyzed the different ontology engineering approaches of each stage of its evolution process, discussing different models of ontology evolution tasks. They suggested the integration of the tools used for ontology evolution and sharing of the research in this field using web portals, besides sharing some common use cases that need to evolve. This survey can help in understanding the ontology evolution process and how they change.

Mihindukulasooriya et al. [53] conducted a quantitative analysis studying the evolutions of the *DBpedia*, *schema.org*, *prov-o*, and *foaf* vocabularies. The authors made some recommendations such as, the need for dividing large ontologies into modules to avoid duplicates when adding new terms and adding provenance information beside the generic metadata when a change is made. Roussakis et al. [63] introduced a framework for analyzing the evolution of LOD datasets. Their framework allows users to identify changes in datasets' versions and make a sophisticated analysis of the evolved data. Palma et al. [61] proposed guidelines to execute the ontology evolution activities, starting from requesting a change to publishing an updating ontology. Their approach covered

two aspects: description of the ontology evolution process and the tasks involved and facilitation of the process using semi-automatic techniques. They explained the ontology evolution tasks of requesting a change, planning the change, implementing the change, and verifying and validating the change. While their methodology involved investigating the evolution process in an abstract form, we analyze the evolution of different vocabularies and how they are changed.

Semantic drift is related to ontology evolution and versioning, indicating the change of meanings in concepts. It aims to identify and measure the changes in ontologies over time. Stavropoulos et al. [69] proposed a hybrid approach to measure and visualize the semantic drift in ontologies. The authors named their approach SemaDrift, involving a collection of tools, methods, and metrics that allow users to measure semantic drift. The authors' hybrid method combines already existing identity-based and morphing-based approaches. The identity-based approach evaluates the stability of concepts with an assumption that the meaning of the concepts is known across ontologies, while the morphing-based approach assesses concepts as their identities are unknown across ontologies. To verify their approach, two real-world scenarios were used, namely digital preservation of art and semantic markup of web services. In our work, we focused on analyzing a vast number of vocabularies in terms of the types of changes and then extracting the changed terms.

So far, few works have discussed changes in vocabularies. Most of the reviewed works study the changes of data in the LOD cloud. Furthermore, some studies have analyzed the evolution of vocabularies from the perspective of ontology engineers when they update them. Finally, the studies have analyzed a limited number of vocabularies, while in our work, we have analyzed 994 vocabularies.

2.3 Reuse of Vocabulary Terms

In this section, we review the works that have analyzed the reuse of vocabulary terms by other vocabularies. First, we present a study that analyzed the reuse of terms in the Internet of Things (IoT) ontologies as well as other works that have described an approach to reusing of terms. Then, we review works that have analyzed the reuse of terms in biomedical vocabularies. Finally, we review a survey for studying the strategies for reusing terms.

Noura et al. [57] identified the most popular ontologies on the IoT to identify the most used terms in this domain. They selected 14 ontologies, and found that 71% of the ontologies reuse less than 18% of the terms defined, and 20% of ontologies are not reused at all. Jiménez-Ruiz et al. [36] described a logic-based approach to reusing terms between ontologies. Their approach specified that the reuse should be safe, i. e., the reused terms are valid (have not been changed/deleted in the source). Furthermore, the reuse should be economic, i. e., only the relevant parts of an ontology are imported.

Ghazvinian et al. [22] studied the overlap between 140 biomedical ontologies. They found more than four million mappings between the concepts. Using those mappings, they analyzed the ontologies, their repositories, and how they could help in ontology design and evaluation. Kamdar et al. [39] published a study regarding the reuse of terms in ontologies in the BioPortal repository. The authors found a term overlap of 25-31%, and the percentage of reused terms was less than 9%. However, none of these studies applied network analysis metrics to the evolved ontologies, as done in this work. Furthermore, they studied the mappings and overlap between ontologies only in the biomedical domain, while we analyze the vocabularies from various domains.

Schaible et al. [65] published a survey of the most preferred strategies for reusing vocabulary terms. The participants, 79 Linked Data experts and practitioners, were asked to rank several LOD modeling strategies. The survey concluded that terms widely used are considered a better approach. Furthermore, the use rate of vocabularies is a more important argument for reuse than the frequency of a single vocabulary term. Their survey can help to understand why some terms are frequently used and some not used at all.

The above studies analyzed vocabularies based on the reuse of terms by other vocabularies. The works focused on a limited number of vocabularies, or on specific domains, especially the biomedical ontologies.

2.4 Use and Adoption of Vocabulary Terms for Modeling Data

The following works studied the use of vocabulary terms by data publishers and the adoption of vocabulary changes in the published data. First, we review the works that analyzed the use of the *schema.org* vocabulary for modeling data. Second, we present the works that analyzed the use of vocabulary terms by analyzing specific datasets or specific domains. Finally, we review a study that published a survey on the quality of the published data and its report on the LOV dataset.

In terms of analyzing the use of structured data on the web, some works focused on *schema.org*. Meusel et al. [52] analyzed its evolution and adoption, comparing the use of *schema.org* terms of over four years by extracting the structured data from the web pages using this vocabulary, from *WebDataCommons* microdata

datasets[5]. They extracted the quads whose object or predicate contained *schema.org*. They found that while not all terms were used, deprecated types and properties were still used, also illustrated in this thesis. Furthermore, they found that publishing new types and properties is preferred over using *schema.org*'s extension mechanism. Thus, the authors focused on analyzing only a single but widely used data schema.

Guha et al. [26] investigated the use of *schema.org* in the structured data of a set of web pages. They analyzed a sample of 10 billion web pages crawled from the Google index and *WebDataCommons* and found that about 31 % of those pages had some *schema.org* elements, and estimated that around 12 million websites are using *schema.org* terms. Furthermore, the authors concluded that the use of *schema.org* is supported by third-party tools, such as *Drupal* and *Wordpress*. In contrast to this work, they did not consider the changes in *schema.org* terms. Additionally, in this work, we are not limited to one vocabulary only.

Some works exploited DyLDO to study the use of vocabularies. Dividino et al. [19] analyzed how the use of vocabulary terms on the LOD cloud have changed over time. They studied the combination of terms that describe a resource but did not investigate whether a vocabulary and its terms have changed. The authors applied their analysis on a dataset of 53 weekly snapshots from the DyLDO dataset, as it is also investigated in this work. Käfer et al. [37] observed the documents retrieved from DyLDO over six months,. They analyzed those documents using different factors, such as their lifespan, availability, and change rate. They also analyzed the RDF content that is frequently changed (triple added or removed). Additionally, they observed how links between documents have evolved over time. While their study is important for various areas such as

[5] `http://webdatacommons.org/`, last accessed: November 28, 2019

caching, link maintenance, and versioning, it does not include information about adopting new and deprecated terms.

Gottron et al. [24] provided an analysis of the LOD schema information by analyzing the BTC 2012 dataset in three different levels. The first level concerned unique subject URIs through studying the dependency relations between the classes and their properties. They found a redundancy between classes and the attached properties. The second level addressed PLDs by dividing the BTC 2012 dataset into individual PLDs. They found that for 20 % of the PLDs, the types can be ignored as the properties perfectly predicted the types. The third level focused on the vocabularies by analyzing how important a vocabulary is for describing the data. They stated that data publishers either applied a strong schematic design or applied a combination of a set of vocabularies to model their data. Hartung et al. [29] proposed a framework to analyze the life science ontologies and their instances. They selected 16 life science ontologies from 2004 to make a comparative evaluation of evolution measures. Cardoso et al. [12] analyzed the impact of ontology evolution on existing annotations. They considered over 66 million annotations from 5,000 biomedical articles and ontologies to support semi-automatic annotation maintenance mechanisms.

Furthermore, some studies analyzed the use of vocabularies. Vandenbussche et al. [72] published a report that describes Linked Open Vocabularies (LOV). It provides statistics about LOV and its capabilities, such as the total number of terms and the top-10 searched types and properties. Rathachari et al. [13] proposed a model that facilitates the understanding of organisms. Their model presents the changes in taxonomic knowledge in RDF form. The proposed model acts as a history tracking system for changing terms but gives no information about how and when the types and properties are used, and which PLDs adopted the changed terms. Zaveri et al. [76] published a survey on the quality of the LOD.

They compared 21 approaches from the Semantic Web data quality assessment, and provided a comprehensive list of metrics and aspects. The authors aimed to provide a clear view and a better understanding of the existing LOD quality assessing approaches. This motivates us to study the impact of vocabulary changes on the published data.

So far, the studies that have analyzed the use of vocabulary terms focus on one or a limited number of vocabularies. Furthermore, LOV provided statistics about a large set of vocabularies, but it did not include information about adopting new terms and which PLD uses which vocabulary.

2.5 Analysis of the Evolution of the Linked Vocabularies and Linked Open Data

In this section, we review several works that have studied how information is propagated in social networks, and the works related to the evolution of the linked vocabularies and LOD cloud. We present studies that analyze large networks and link prediction in them. Afterwards, we provide an overview of the LOV dataset, which represents a set of linked vocabularies, then review the studies regarding analyzing the evolution of the LOD cloud on the instance level. Subsequently, we present the works that have analyzed the network of linked vocabularies based on a specific domain such as tourism. Finally, we review the works that have analyzed the changes in RDF documents.

Teng et al. [70] studied information entropy encoded in social networks and information novelty. In this context, influence mining aims to detect indirect influences between two nodes in a network for tasks like marketing and

recommendation [33, 50]. Hu and Cao [32] proposed a probabilistic model to analyze the influences in networks like retweets on Twitter. Ohsaka et al. [59] computed the k most influential nodes as well as nodes that may be influenced if some other nodes are activated. Besides the network structure, textual content associated with the nodes, like tweets on Twitter, are also used for network analysis [4, 10]. Kaur and Singh [41] published a survey for reviewing different data mining approaches for detecting anomalies in social networks. The authors defined the term anomaly as abnormal behavior compared to others on the same network structure. They employed supervised, semi-supervised and unsupervised anomaly detection methods.

Other works focus on link prediction in networks to determine between which nodes a new edge may appear [64, 48]. Leskovec and Sosič [49] presented a platform that provides high-level operations to analyze large networks, namely the Stanford Network Analysis Platform (SNAP), which can efficiently add or remove nodes and edges from/to the network. Furthermore, SNAP needs a smaller amount of memory, compared with other network analytic systems, such as NetworkX [27] and iGraph [15]. Additionally, SNAP is fully dynamic when dealing with large graphs, i. e., the network's structure and attributes can be modified during the computation.

Khan et al. [42] provided a tool to facilitate the visualization of large networks. They discussed different graph summarization algorithms for both static and dynamic networks. The authors discussed four summarizing techniques, namely aggregation-based techniques, attribute-based techniques, compression methods, and application-oriented techniques. Furthermore, the authors compared the different summarization techniques based on the metrics of space requirement, efficiency, accuracy, and interest level. Dietz et al. [16] visualized research areas

extracted from a citation network to describe the flow of topics between papers and assess the impact of papers.

Vandenbussche et al. [72] published a report that describes LOV and provides some descriptive statistics. They also provided a system that shows the dependencies between vocabularies, but it does not give information about the imported types and properties. Furthermore, there is a lack of information about the statistics regarding the reuse of terms in the network of vocabularies, such as the most reused terms and whether deleted types and properties are still reused.

Vassilis et al. [62] proposed a benchmark generator to evaluate the ability of the current versioning strategies to manage LOD datasets. Their benchmark generator can produce different sizes of data and apply a managed number of insertion and deletion actions for different versions of the generated data, as well as the SPARQL queries. The authors tested their benchmark using R43PLES (revision of triples). Akhtar et al. [5] proposed an approach to update the local caches by identifying the important changes in the LOD cloud, by capturing the changes and giving them weights. To update the caches, the authors proposed an updating strategy by combining the estimated changes with the current local copy of data. The authors evaluated their approach based on the F1 score using the BTC and DyLDO datasets. Compared with the existing updating approaches, their results showed that the accuracy was 88%, the precision score ranged between 0.883 and 0.890, and the recall score ranged from 0.884 to 0.894.

Umbrich et al. [71] applied *k*-means clustering to find groups of data items with similar changes in randomly sampled data sources. Manual inspection revealed that data items from the same domain often share similar temporal characteristics. However, the authors only considered whether there was a change and did not take into account the number of statements that changed. Furthermore, Umbrich

et al. [71] collected 24 weekly snapshots of the neighbors of a single seed URI, the data profile of Tim Berners-Lee. During this time, 35% of the RDF documents had changed.

Käfer et al. [37] collected 29 weekly snapshots of a seed list with 86,696 RDF documents and analyzed the changes between pairs of two consecutive snapshots; the results showed that RDF documents change frequently. Gottron and Gottron [23] compared the accuracy of various RDF indices over the weekly snapshots from Käfer et al. [37]. Dividino et al. [17] analyzed the dynamics of the data by Käfer et al. [37] and proposed a monotone, non-negative function to represent the dynamics of RDF statements as single numerical value. They also developed an adaptive crawling strategy to keep caches of RDF data up-to-date while respecting limited bandwidth for crawling. Nishioka and Scherp [55] computed periodicities of temporal changes in the dataset by Käfer et al. [37], and conducted information-theoretic analyses of the data evolution [56].

Few studies have been conducted on analyzing the evolution of linked vocabularies. Furthermore, most of the analyses of LOD focused on the instance level. Additionally, some works focused only on a specific domain without expanding these analyses to include multiple domains.

2.6 Summary

In this section, we introduced the Semantic Web principles and terminology, and summarized the works related to vocabulary changes and evolution. We presented the limitations regarding the changes of vocabularies and reuse of terms, the use

of vocabularies types and properties by data publishers, and the evolution of the NeLO and the LOD cloud. The key differences from our work are as follows:

- Regarding the vocabulary changes and the reuse of terms between vocabularies, the current literature focused on analyzing one or few vocabularies and did not investigate how the studied vocabularies are changed. Furthermore, most of the studies focus on a specific domain. Therefore, there is a lack in studies that include most, if not all the domains of vocabularies. We provide an analysis for a wider range of vocabularies and show how they have changed. Additionally, we provide an in-depth study on how other vocabularies reuse the vocabulary terms and give detailed results.

- For the adoption of vocabulary types and properties and the reuse of those terms by the data publishers, many works in the literature analyzed the use of vocabulary terms on the LOD cloud. Those works are essential, but they do not include information about adopting new and deprecated terms. Furthermore, other works provided detailed statistics about the vocabularies in the LOD cloud. Those studies are useful, but they did not give answers about how and when the types and properties are used, and which PLDs adopted the changed terms.

- Regarding the NeLO, few works have been done in this area. Most of the works focus on the evolution of the LOD cloud. Furthermore, there is a lack of information regarding the network's evolution, such as the most reused terms and whether the terms that have been deleted are still reused. Thus, we consider the schema-level, i. e., the evolution of vocabulary terms reused in other vocabularies, and the impact of the vocabulary changes on other vocabularies. Furthermore, we provide diagrams which visualize the

evolution and statistics for the evolution of the NeLO over time, from 2001 until 2018.

Chapter 3

Vocabulary Changes and Reuse

Vocabularies are subject to change during their lifetime. Changes are required to reflect new requirements, shifts in the domains of the vocabularies model, or handling errors that have appeared in prior versions [44]. Since the reuse of types and properties between vocabularies is one of the main principles of the Semantic Web, the changes in a vocabulary must be reflected by the other vocabularies that reuse these changed terms.

In this chapter, we provide a detailed analysis of the changes in vocabularies. We analyze the changes in the types and properties for a set of vocabularies, and the reuse of those terms in other vocabularies. We use two steps: First, we analyze Schmachtenberg et al.'s [67] state of the LOD cloud report to select the 12 most dominant vocabularies in terms of their use, in different pay level domains. Subsequently, we used the LOV dataset[1] to download the available version of those 12 vocabularies. The number of versions we found for these vocabularies

[1] `http://lov.okfn.org/dataset/lov/`, last accessed: November 28, 2019

range between two to 11. While some vocabularies exist for more than 10 years (e. g., *foaf*), others are only online for around two years (such as *dcat*).

To analyze the reuse of vocabulary terms by other vocabularies, we extracted all types and properties from the vocabularies listed on LOV, and their different versions, until June 2018. We examined 636 vocabularies listed in LOV. The extracted types and properties have been classified into two categories: the *own terms* are the terms created by the ontology engineers of the considered vocabulary, while the *imported terms* are reused from other vocabularies.

The methodology we used to analyze the change of vocabularies and the reuse of terms by other vocabularies are described in Section 3.1. We describe the methodology used to analyze the changes of the 12 most dominant vocabularies in the state of the LOD cloud report in Section 3.1.1. We describe the process of analyzing the 636 vocabularies in the LOV dataset in Section 3.1.2. The datasets used in the experiments are briefly described in Section 3.2, and the results are shown in Section 3.3. In Section 3.4, we discuss the findings of analyzing the changes of vocabularies and the reuse of types and properties by other vocabularies, and conclude the analysis in Section 3.5.

3.1 Analysis Methodology

The methodology for analyzing our selected set of vocabularies is shown in Section 3.1.1. The vocabularies are the most dominant ones, based on the state of the LOD cloud 2014 report [67]. Subsequently, in Section 3.1.2, we extend the methodology in Section 3.1.1 to include more vocabularies, and analyze the reuse of vocabulary types and properties by other vocabularies.

3.1.1 Analysis Methodology of Changes of Vocabularies

In April 2014, Schmachtenberg et al. [67] published a report providing detailed statistics about the LOD cloud. The authors analyzed a subset of data from the LOD cloud, based on crawling seed URIs from the datahub.io[2] dataset, the BTC 2012 dataset[3], and the public-lod@w3.org[4] mailing list.

Based on Schmachtenberg et al.'s report, we selected the most dominant vocabularies in their crawled subset of the LOD cloud. The methodology can be expressed as follows. First, to analyze changes in the vocabulary, we need to have at least two available versions of the considered vocabulary to download. Second, we chose the vocabularies that have been used in more than five LOD datasets (0.49% of 1,014 datasets used in the statistics of the state of the LOD cloud report). We excluded the vocabularies that have been used in less than five datasets because they are rarely used. The excluded vocabularies constituting 50% of the vocabularies listed in the state of the LOD cloud 2014 report. Furthermore, we excluded the *owl* meta-vocabulary, because of all our selected set of vocabularies, in all their versions, reuse the same terms from *owl*. The reused terms are five annotation properties (`backwardCompatibleWith`, `deprecated`, `incompatibleWith`, `priorVersion`, and `versionInfo`) as well as "`Thing`" type. Therefore, we decided not to include it in this part of the work nor the previous meta-vocabularies [1].

Based on our criteria for selecting the vocabularies, we obtained 62 vocabularies that had been used in more than five datasets. We found 12 vocabularies from the 62 with more than one version and that could be downloaded. The 12 vocabularies were related to all topical domains in the LOD cloud, which are

[2]http://datahub.io/dataset?tags=lod
[3]http://km.aifb.kit.edu/projects/btc-2012/
[4]http://lists.w3.org/Archives/Public/public-lod/

government, publications, life sciences, user-generated content, cross-domain, media, geographic, and social web. The topical domains are the domains of the datasets that the published data represent.

Afterwards, we downloaded all available versions for the 12 vocabularies from the LOV dataset and the official websites of some vocabularies that we could not find on the LOV dataset. By using Protégé 4.3.0[5], we extracted the differences between every two successive versions to capture the changes of those vocabularies. Table 3.1 shows the number of downloaded versions for each vocabulary, the period from the first to the latest version for those vocabularies, the evolution duration in years/months, and the average number of changes per year.

Subsequently, we analyzed the changes that we captured in different versions of the vocabularies. The changes are classified into two types: create, and delete. These changes can be in terms of types (classes), object properties, data properties, annotation properties, data types, or individuals. We analyzed the changes using different aspects. First, we counted the number of changes for each type of terms, i.e., types and properties. Afterwards, we observed the percentage of internal changes versus external changes. Internal changes are changes that occurred in the types and properties that were initially introduced and developed by ontology engineers of the vocabulary. External changes are changes in the vocabularies that export terms to the vocabulary.

[5]http://protege.stanford.edu

Table 3.1: The number of downloaded versions of the examined vocabularies, sorted by the number of datasets they were used based on the state of the LOD cloud report 2014. The table shows the evolution period in years/months and the average number of changes per year.

Vocabulary	#Versions	Duration in years/months	Average number of changes/year
foaf[6]	10	8 years & 10 months	30
dcterms[7]	3	4 years & 5 months	26
skos[8]	8	5 years & 4 months	44
cube[9]	4	3 years & 8 months	10
bibo[10]	2	1 year & 5 months	13
dcat[11]	6	2 years & 1 month	47
gn[12]	11	6 years & 1 month	34
swc[13]	3	3 months	184
prov[14]	3	2 years & 8 months	106
aiiso[15]	3	4 months	79
org[16]	10	3 years & 10 months	70
cal[17]	2	10 years & 9 months	6

3.1.2 Analysis Methodology of the Reuse of Terms by other Vocabularies

To analyze the reuse of vocabulary types and properties by other vocabularies, we expanded the selection criteria described in Section 3.1.1 to include all the vocabularies in the LOV. We designed a methodology consisting the following two steps:

1. We extracted all types and properties from all available versions of vocabularies (from June 2001 until June 2018) listed in the LOV dataset. The types and properties extracted were classified into two categories: the

[6] `http://www.foaf-project.org/`, last accessed: November 28, 2019
[7] `http://dublincore.org/documents/dcmi-terms`, last accessed: November 28, 2019
[8] `http://www.w3.org/2009/08/skos-reference/skos.html`, last accessed: November 28, 2019
[9] `http://www.w3.org/TR/vocab-data-cube/`, last accessed: November 28, 2019
[10] `http://bibliontology.com/`, last accessed: November 28, 2019
[11] `https://www.w3.org/TR/vocab-dcat/`, last accessed: November 28, 2019
[12] `http://www.geonames.org/ontology/documentation.html`, last accessed: November 28, 2019
[13] `http://data.semanticweb.org/ns/swc/swc_2009-05-09.html`, last accessed: November 28, 2019
[14] `https://www.w3.org/TR/prov-o/`, last accessed: November 28, 2019
[15] `http://vocab.org/aiiso/schema-20080925.html`, last accessed: November 28, 2019
[16] `https://www.w3.org/TR/vocab-org/`, last accessed: November 28, 2019
[17] `http://www.w3.org/TR/rdfcal/`, last accessed: November 28, 2019

own terms are terms originally created by the ontology engineers of the considered vocabulary V. The *imported terms* are the terms reused by vocabulary V via reuse from other vocabularies.

2. We checked whether the *imported terms* were the most recent ones, i. e., if the types and properties that appear in the latest published version of the source vocabulary are actually those that are reused in the target vocabulary or if older versions of terms are considered.

For the first step, we examined 636 vocabularies listed in LOV. We employed the OWL API[18] version 5.1.6 to extract all the *own terms* and *reused terms* from the latest version of all the 636 vocabularies. The OWL API is an open-source Java API used to create, manipulate, and parse ontologies. While extracting the reused terms, some additional vocabularies not contained in LOV were found [3]. Thus, we considered a total of 994 vocabularies.

For the second step of the methodology, we found all the deleted and deprecated terms of the vocabularies by parsing and comparing all versions of vocabulary and recording the types and properties that were deleted or deprecated during the lifespan of each vocabulary.

3.2 Datasets

State of the LOD cloud report 2014: It is a report about the structure and content of a large-scale snapshot of the LOD cloud [68], published by Schmachtenberg et al. in 2014. To crawl that snapshot, they used the

[18] `https://github.com/owlcs/owlapi`, last accessed: November 28, 2019

LDSpider [34], which is a framework to harvest linked data. They used 560 thousand URIs as a seed list for the crawler. The report shows the relationships between the LOD datasets. The authors found that the links between datasets had doubled between 2011 and 2014.

Linked Open Vocabularies (LOV): LOV is a dataset consisting of *rdfs* and *owl* vocabularies from the LOD cloud [72]. LOV uses a script that automatically checks for vocabulary changes on a daily basis and stores the detected version locally. It provides a set of features such as vocabulary documentation, access to LOV code and data, and a search engine for vocabularies. Furthermore, it provides a history for the prior versions of vocabularies which are downloadable.

3.3 Results

In this section, we summarize the findings based on the experiments conducted in Sections 3.1.1 and 3.1.2. In Section 3.3.1, we list our findings based on the methodology explained in Section 3.1.1 [1]. The results generated using the experiments explained using the methodology in Section 3.1.2 are listed in Section 3.3.2 [3].

3.3.1 Changes of Vocabularies

Table 3.2 shows the domains for the selected vocabularies based on the methodology described in Section 3.1.1. Additionally, the table shows the percentage of the datasets that use these vocabularies, based on the results

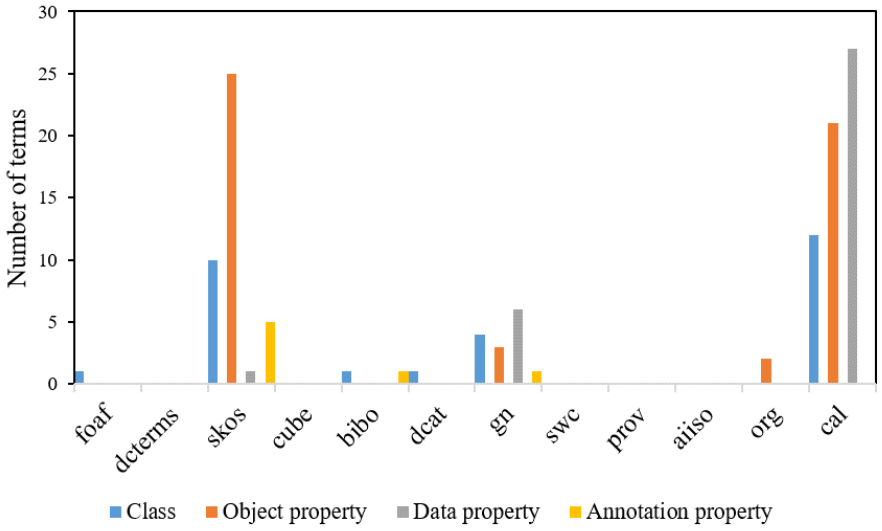
Table 3.2: Vocabularies according to their domains and the percentage of dataset they appear in based on the state of the Linked Open Data cloud report 2014.

Vocabulary	Domain	Number of datasets
foaf	Cross-domain	701 (69.13%)
dcterms	Cross-domain	568 (56.02%)
skos	Publications/ Cross-domain/ Geographic	143 (14.10%)
cube	Government/ Geographic	114 (11.24%)
bibo	Cross-domain/ Social web/ Media/Publications/Life Sciences	62 (6.11%)
dcat	User-generated content/Cross-domain	59 (5.82%)
gn	Geographic/ Life Sciences/ Media/ Social web	27 (2.66%)
swc	Social web	27 (2.66%)
prov	Government/ Cross-domain	21 (2.07%)
aiiso	Publications/ Life Sciences	17 (1.68%)
org	Social web	14 (1.38%)
cal	Social web	9 (0.89%)

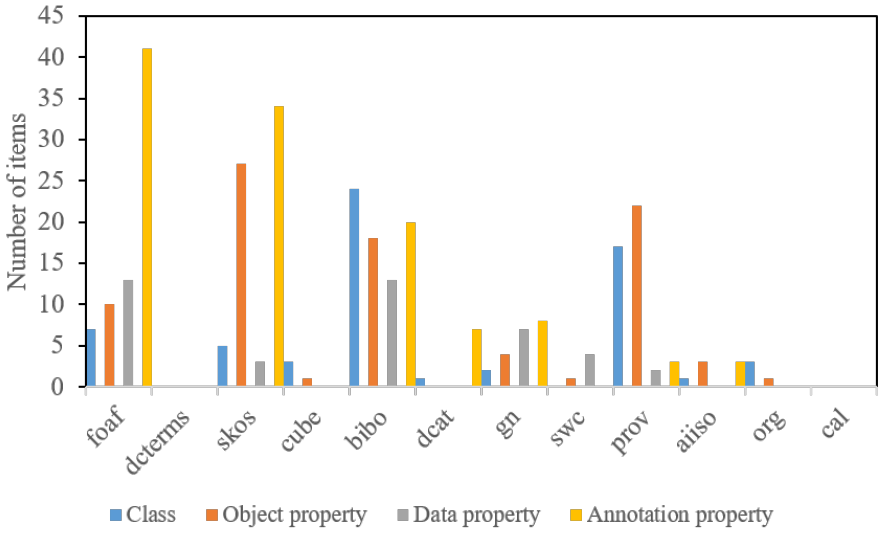
published in the state of the LOD cloud report. The table shows 12 vocabularies. Please note that the selected vocabularies are based on the availability of versions (if they had any). Any vocabulary that had only one version, or its previous versions were unavailable to download, were excluded from this analysis. Based on the history information from the LOV, we found that 65% of the 62 examined vocabularies had one version. Furthermore, we found that 15% of those 62 vocabularies had more than one version, but were unavailable to download. Thus, we excluded them from this study. In Table 3.2, we can see that more than 50% of the datasets of the state of the LOD cloud report use two vocabularies, namely *foaf* and *dcterms*.

Figure 3.1 represents the total number of each type of change. The change can occur in classes, object properties, data properties, or annotation properties. From Figure 3.1a, we can observe that most of the changes are related to the "creation" type, mostly in the annotation properties.

The percentage of changes of internal terms compared to external terms in the vocabulary versions are shown in Table 3.3. We calculated the total percentage for all internal changes through out the vocabulary evolution period, i. e., all versions. We can see that more than 90% of the changes of *dcterms*, *gn*, and *prov* occurred



(a) Total number of created items



(b) Total number of deleted items

Figure 3.1: The total number of the changed classes, object properties, data properties, and annotation properties, for each type of change.

Table 3.3: The percentage of the internal changes compared to the total changes.

Vocabulary	Percentage of the internal changes compared to the total changes
dcterms	99%
prov	98%
gn	97%
skos	89%
aiiso	89%
swc	86%
foaf	79%
cal	77%
org	73%
dcat	71%
bibo	65%
cube	43%

for the internal terms. Furthermore, we can see that *cube* is the only vocabulary with an internal changes percentage less than 50%.

Table 3.4 shows the imported vocabularies that are reused for creating each considered vocabulary. The "*Version*" column shows in which version the considered vocabulary imported the vocabularies specified in the "*Imported vocabularies*" column. From the imported vocabularies listed in this table, we can notice two things: First, there are three vocabularies, *foaf*, *dcterms* and *swc*, that kept their imported vocabularies unchanged from the first version until the latest. Second, we can note that there are two vocabularies (*gn* and *org*) where many changes in the vocabularies they import. Over six years, *gn* partially changed the imported vocabularies five times, and in approximately four years, *org* changed the imported vocabularies four times.

Table 3.4: Vocabularies in LOV, the vocabularies imported, and the version.

Vocab.	Imported vocabularies	Version
foaf	dc/owl/rdf/rdfs/vs/wot/xml/xsd	All versions
dcterms	dcam/owl/rdf/rdfs/skos/xml/xsd	All versions
skos	dc/dcterms/owl/rdf/rdfs	First version
	dc/dcterms/owl/rdf/rdfs/foaf/vs	Mar 2005-Aug 2008
	dcterms/owl/rdf/rdfs	Mar 2009-Aug 2009
cube	dcterms/foaf/owl/rdf/rdfs/scovo/skos/void/xml/xsd	All versions
bibo	Address/dc/dcterms/event/foaf/owl/rdf/rdfs/skos/time/vann/vs/wgs84_pos/xml/xsd	Jun 2008
	Event/foaf/ns/owl prism/rdf/rdfs/schema/skos/dcterms/vann/xml/xsd	Nov 2009
dcat	dcterms/owl/rdf/rdfs/xml/xsd	First version
	dc/dcterms/dctype/foaf/owl/rdf/rdfs/schema/skos/vcard/vann/voaf/xml/xsd	Remaining versions
gn	skos/owl/rdf/rdfs/xml/xsd	Oct 2006-May 2010
	cc/dcterms/foaf/owl/rdf/rdfs/skos/wg84_pos/xml/xsd	Sep 2010
	dcterms/foaf/owl/rdf/rdfs/skos/xml/xsd	Oct 2010
	cc/dcterms/foaf/owl/rdf/rdfs/skos/vann/voaf/xml/xsd	Feb 2012
	adms/cc/dcterms/foaf/mrel/owl/rdf/rdfs/skos/vann/xml/xsd	Oct 2012
swc	bibtex/dc/dcterms/doap/foaf/geo/cal/misc/owl/rdf/rdfs/sioc/swrc_ext/vcard/vs/wordnet/xml/xsd	All versions
prov	owl/rdf/rdfs/skos/xml/xsd	First version
	owl/rdf/rdfs/xml/xsd	Remaining versions
aiiso	cc/dc/dcterms/dctype/owl/rdf/rdfs/skos/vann/xml/xsd	In first two versions
	cc/dc/dcterms/dctype/foaf/owl/rdf/rdfs/skos/vann/vs/xml/xsd	In latest version
org	dc/foaf/gr/opmv/owl/time/rdf/rdfs/skos/vcard/xml/xsd	In first version
	dcterms/foaf/gr/opmv/owl/time/rdf/rdfs/skos/vcard/xml/xsd	Oct 2010-Sep 2012
	dcterms/foaf/gr/opmv/owl/time/prov/rdf/rdfs/skos/vcard/xml/xsd	Oct 2012
	dcterms/foaf/gr/owl/time/prov/rdf/rdfs/skos/vcard/xml/xsd	Feb 2012-Apr 2014
cal	dt/owl/rdf/rdfs/xml/xsd	First version
	dc/dt/xhtml/owl/rdf/rdfs/xml/xsd	Latest version

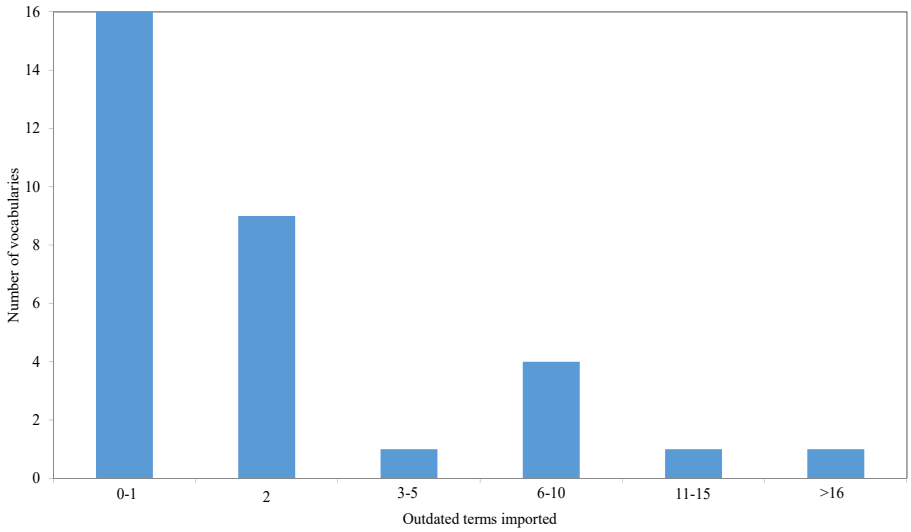


Figure 3.2: The number of vocabularies that import outdated terms aggregated by the number of outdated terms imported.

3.3.2 Reuse of Terms by other Vocabularies

Figure 3.2 shows a histogram of the vocabularies from the LOV that reuse outdated types and properties (Section 2.1) from other vocabularies. On the one hand, we can see that 16 vocabularies import only one outdated term. On the other hand, six vocabularies import more than six outdated types or properties. Furthermore, 10 vocabularies import between two and five outdated terms.

Three vocabularies removed the imported terms after they were deleted from their original vocabularies, as listed in Table 3.5 The "*Updated version*" column represents the date the vocabulary was updated, and the "*Prior version*" contains the terms before they became outdated. Notably, the *oslo* vocabulary removed five outdated terms, but it still reuses two outdated terms in its latest version.

Since reusing existing terms is one of the main principles of linked data, we show

Table 3.5: Vocabularies that removed the outdated types and properties from prior versions.

Vocabulary	# outdated terms removed/ # outdated terms imported	Updated version	Prior version
qudt	12 / 12	9-Oct-2016	1-Jun-2011
oslo	5 / 7	30-May-2014	30-Sep-2013
dcat	1 / 1	28-Nov-2013	20-Sep-2013

Table 3.6: Top-10 types and properties that are imported by other vocabularies listed on Linked Open Vocabulary (LOV).

Term	Importing vocabularies
`dcterms:modified`	281
`dcterms:title`	276
`dce:title`	266
`dce:creator`	263
`vann:preferredNamespacePrefix`	257
`dcterms:description`	249
`vann:preferredNamespaceUri`	241
`foaf:Person`	175
`foaf:name`	164
`cc:license`	122

in Table 3.6 the top-10 terms imported by the vocabularies in the LOD cloud. Please note, we removed meta-vocabularies (Section 2.1) from this table since it is quite natural that they are mostly used. The terms, i. e., types and properties, were extracted from the latest versions of vocabularies listed in the LOV. We can see that `dcterms:modified` is the most imported term, and defined in the *dcterms* vocabulary, and its property is imported by 257 vocabularies.

3.4 Discussion

Most of the vocabularies are highly static, and most of the changes that occurred were related on the terms created by the ontology engineers of the changed vocabulary (internal changes). Furthermore, we found that not all vocabularies

reuse the most recent types and properties from the vocabularies they import. In Section 3.4.1, we discuss the results regarding the changes in vocabularies. The results related to the reuse of types and properties of the vocabularies of the LOV dataset are discussed in Section 3.4.2

3.4.1 Changes of Vocabularies

The majority of changes occurred in the annotation properties. An annotation property is used to provide for more explanations (metadata) to clarify types, properties, or individuals. Another observation is that vocabularies are highly static with respect to the number of the imported vocabularies [1]. Based on the low number of changes, we can conclude that the domains are almost entirely covered with types and properties in the existing vocabularies.

In most of the examined vocabularies, we found that most of the changed types and properties occurred on the internal terms. On the other hand, some vocabularies, such as *cube*, *bibo*, *dcat*, and *org*, were changed because of changes in their external vocabularies. For example, in the *cube* vocabulary, the percentage of the internal change is 43%, and the external terms cause the remaining 57% of changes that *cube* imports. Another example is the *bibo* vocabulary. Analyzing 10 versions for this vocabulary, we conclude that external vocabularies cause 35% of changes. The *bibo* vocabulary uses many external vocabularies such as *dcterms* and *skos*. Both of those external vocabularies have versions published within the two published versions of *bibo*, the first in June 2008, and the latest version in November 2009. We can conclude that there is always a need to keep track of the changed item imported from other vocabularies.

Vocabularies such as *dcterms*, *gn*, and *prov* keep updating the types and properties that have been created. For example, we analyzed 11 versions of the *gn* vocabulary, and the percentage of internal changes was 97%. Thus, when the ontology engineers need to make a change, they change the terms. Another observation regarding the *dcterms* and *prov* vocabularies is that they imported only a few vocabularies, as shown in Table 3.3, and most of them in those external vocabularies are meta-vocabularies.

The vocabularies change over the lifespan. Some vocabularies such as *foaf*, *skos*, *gn*, and *org* have many versions; 10, 8, 11, and 10, respectively. The *gn* vocabulary had 11 versions in the period from October 2006 to October 2012, and *org* has 10 versions in the period from June 2010 to April 2014. We think that this large number is caused by the domain they are related to, especially the social web domain that has shown fast growth in the last few years, to reflect the shifts in their domains. For example, *foaf* has published 10 versions in approximately nine years.

New versions of vocabularies, together with the great variety of vocabularies already existing, and the new vocabularies, may overwhelm ontology engineers. This allows them to choose from a vast amount of alternative terms when building or updating their ontologies. Similar issues may occur for data publishers when deciding which vocabularies to reuse for modeling their datasets. Missing some changes and consequently not updating an ontology or a dataset is likely (see Section 3.4.2), notably in a distributed environment such as the LOD cloud. This holds particularly for deprecation and deletion, where these types of change are more critical.

3.4.2 Reuse of Terms by other Vocabularies

Many vocabularies are up-to-date with respect to to their imported terms. 35 vocabularies were affected by term updates in other vocabularies. We found that 33 vocabularies reuse outdated types or properties. The remaining two vocabularies (*qudt* and *dcat*) removed the terms they imported when they become outdated. Although this number of outdated vocabularies may seem low, it can have a substantial impact on the published data. The number of outdated terms reused by those vocabularies ranges between one and 20. For example, the *dpn* vocabulary reuses 20 outdated terms, and the *voag* vocabulary reuses 15 outdated terms [3]. Furthermore, there are 16 vocabularies that reuse only one outdated term (please refer to Chapter 4 for more details on the impact of vocabulary changes on the published data in the LOD cloud). We think that the process of checking for changes to update the ontologies is done manually, since we found that the vocabularies excluded some outdated terms while keeping others.

Of the 35 vocabularies affected by changes, three have been updated by removing some of the outdated terms. For instance, the *oslo* vocabulary removed five terms, one from *adms* and four from *rov*. However, *oslo* still reuses two terms from *vcard*, although they have been deleted in *vcard*. This could either mean that the deleted terms are still needed and no alternatives have been found, or that some updates have been missed because the process for looking for changes is done manually. Reusing terms from older vocabulary versions, which can still be accessed by the URI of the version, is possible, but we recommend checking the reasons for deleting such terms.

Overall, 16% of the terms of the vocabularies listed in LOV are reused by other vocabularies. This number is still low, and there is a need to increase the reuse of the existing types and properties in order to avoid overlap and redundancy in

the data representation [35]. Tools to suggest existing types and properties like TermPicker [66] can play a major role in increasing the number of terms reused by helping ontology engineers to select and discover the existing terms to reuse.

The SemWeb Vocabulary Status (*vs*)[19] ontology provides information about the status of a term, but it is not widely used. This vocabulary (or similar ones) can help ontology engineers to check the recent status of terms before reusing them, i. e. to avoid reusing unstable terms that are likely to be removed in the future.

Tools that notify ontology engineers about vocabulary changes may help them track the changes and reduce the update effort. Tool support becomes especially important when a vocabulary has many dependencies: the more terms the ontology reuses, the higher is the effort to update the vocabulary when a change occurs. With many dependencies, it is challenging to keep an ontology up-to-date, as any change in one of the imported vocabularies could require an update of the importer. Some vocabularies import terms from more than 40 other vocabularies. Overall, 12% of the vocabularies imported from 59% of the other vocabularies, accounting for 22% of incoming imports.

3.5 Summary

We analyzed the changes in vocabularies on the LOD cloud and the reuse of terms by other vocabularies (RQ1 Section 1.2). Based on the statistics of the LOD cloud 2014 report, we selected the twelve most dominant vocabularies in terms of their use in different pay level domains. The number of versions we found for these vocabularies ranged between two to 11. While some vocabularies exist for more

[19] `https://www.w3.org/2003/06/sw-vocab-status/note`

than 10 years (e.g., *foaf*), others have only been online for a few years (like *dcat*). Furthermore, we analyzed a broad set of vocabularies (626 vocabularies) from the LOV dataset to investigate the reuse of vocabulary types and properties between vocabularies, and how the change in one vocabulary can affect others who import it.

Most of the vocabularies are highly static, and most of the changes occurred in the annotation properties of the vocabularies, which are designed to add clarity to types and properties defined in the vocabulary. When a change is needed, most of the changed types and properties are the ones created by the ontology engineers of the considered vocabulary, i. e., own terms. There are some exceptions for this observation, but most of the changed vocabularies follow this behavior. Most of them show increases in their number of types and properties, which leads to more knowledge represented in the LOD cloud. The great number of vocabularies for each topical domain, and the continued growth of vocabularies can overwhelm ontology engineers when they update their vocabularies. Therefore, there is a need to regularly check for changes in the vocabularies to keep them and the datasets updated. Regarding the reused types and properties of the vocabularies listed in the LOV, only 16% of the existing terms are reused by the other vocabularies, which can be considered low. Using recommender systems can help increase the reuse amount between vocabularies in order to avoid overlap and redundancy. Usually, the process of checking for changes in vocabularies is done manually. We observed that 33 vocabularies in the LOD cloud still reuse deprecated or deleted terms. There is a need for tools that notify ontology engineers and data publishers about the changes of the vocabularies. Knowing what has changed can help update the vocabularies regularly, rather than checking for updates manually, which may lead to missing some updates.

Chapter 4

Use and Adoption of Vocabulary Terms for Modeling Data

In this chapter, we present our analysis of the use and adoption of vocabulary terms to better understand by whom and how the vocabularies are used, and how vocabulary changes are adopted in data by data publishers.

Most of the newly coined terms are adopted in less than one week after their publishing date. However, some terms are only adopted after several months or even years after the date of creation, while some others appear even before their official publishing date. Many deprecated and deleted terms are still in use in data; therefore, those terms are not really deprecated or deleted. For most vocabularies, notably in the BTC dataset, more than 50 % of terms are actually unused. We found no deprecation of terms in the Wikidata vocabulary. Moreover, 17 terms are not used because they are for defining properties and their types.

In Section 4.1, we describe the analysis methodology of selecting vocabularies with multiple versions to analyze their changes and see how they are adopted in the data. In Section 4.2, we briefly describe the datasets that were applied. We used three well-known datasets that crawl data from the LOD cloud: DyLDO [38], BTC[1], and Wikidata[2]. The first dataset is a collection of weekly snapshots for a set of linked data documents from the LOD cloud. The second dataset was formed by yearly crawling from the LOD cloud from 2009 to 2012, as well as in 2014. From both datasets, we extracted the PLDs using terms from a selected set of vocabularies and adopted the changes of vocabulary terms. For the last dataset, we extracted the vocabulary types and properties from Wikidata and determined whether changes on the vocabulary were done (additions or deprecations/deletions) and how these changes were adopted in the Wikidata dataset. The results of the analysis are presented in Section 4.3. In Section 4.4, we discuss the findings of analyzing the use of vocabulary terms by data publishers and the adoption of their changes, before summarizing.

4.1 Analysis Methodology

To analyze the use of vocabulary terms and adoption of the newly created terms in data, we followed two steps. First, we determined the vocabularies that have more than one published version on the web to extract the changes between the successive versions and analyzed the adoption of these changes. Second, we investigated how data publishers use vocabulary types and properties as well as how the changed terms of vocabularies are adopted and used for modeling data by parsing three well-known datasets.

[1] `http://challenge.semanticweb.org/`, last accessed: November 28, 2019
[2] `https://www.wikidata.org`, last accessed: November 28, 2019

For the first step, we relied on Schmachtenberg et al.'s [68] published report with detailed statistics about a large-scale snapshot of the LOD cloud. The snapshot comprises seed URIs from the datahub.io dataset[3], the BTC 2012 dataset[4], and the public-lod@w3.org mailing list[5]. We selected a set of vocabularies that satisfy the following conditions and characteristics for the analysis.

1. The vocabulary has at least two versions published on the web to enable the capturing and extracting of the changes. These versions must be available to download and use.

2. At least one version is covered by the datasets under investigation. For example, for the DyLDO dataset, there needed to be one version of the vocabularies published after May 6, 2012, since that is the first snapshot of the DyLDO dataset was crawled.

3. The vocabulary terms are directly used for modeling some data, i. e., they occur in at least one triple in the published dataset. In contrast, vocabularies could also just be linked by a data publisher, where changes of external vocabularies may not have any impact on the published data.

On the basis of these criteria, we obtained 134 of the most dominant vocabularies listed in the state of the LOD cloud 2014 report by Schmachtenberg et al. [67]. We found 18 vocabularies with more than one version. From them, 13 vocabularies had changes (either additions or deprecations/deletions) in terms created by the ontology engineers (own terms), in the timeframe of the considered datasets. The remaining five vocabularies had changes on the terms they imported from other vocabularies, not on their own terms [2]. We downloaded the different versions

[3] `http://datahub.io/group/lodcloud`, last accessed: November 28, 2019
[4] `http://km.aifb.kit.edu/projects/btc-2012/`, last accessed: November 28, 2019
[5] `http://lists.w3.org/Archives/Public/public-lod/`, last accessed: November 28, 2019

Table 4.1: Overview of the vocabularies, the available versions, and the number of changes.

Vocabulary	Versions	Changes
adms[8]	2	18
cito[9]	3	218
cube[10]	2	6
dcat[11]	2	13
emp[12]	2	1
geom[13]	2	2
gn[14]	7	31
mo[15]	2	46
oa[16]	2	31
org[17]	2	8
prov[18]	5	168
voaf[19]	4	8
xkos[20]	2	1

using the LOV observatory[6]. We exploited the PromptDiff Protégé 4.3.0[7] plugin to identify the vocabulary changes. PromptDiff is an ontology versioning tool that extracts the changes between two versions of ontologies [58], by identifying the changes and showing the difference between two versions.

For the second step of the methodology, we analyzed the vocabularies resulted from the first step. We analyzed them in terms of their use in the LOD cloud and Wikidata, and the adoption of the vocabulary changes in those datasets. The resulted vocabularies are listed in Table 4.1. The table also provides the number of versions considered for each vocabulary, and the total number of changes (additions and deletions) [2].

[6] `http://lov.okfn.org/dataset/lov`, last accessed: November 28, 2019
[7] `http://protege.stanford.edu`, last accessed: November 28, 2019

4.2 Datasets

We analyzed three well-known, large-scale datasets with reference to their use of the vocabularies. The first datasets are DyLDO and BTC, which were obtained from the LOD cloud, and the third is Wikidata. We analyzed the use of vocabulary terms, and adoption of vocabulary changes within each dataset. We used the LOV to download all available versions for the resulted vocabularies, after applying the methodology in Section 4.1. Below, we briefly summarize the main characteristics of each of the three datasets. Furthermore, we describe the LOV dataset, which we used to download the versions of vocabularies used in this analysis.

The Dynamic Linked Data Observatory (DyLDO): DyLDO is a repository to store weekly snapshots from a subset of web data documents [38]. The main idea is to collect frequent, continuous snapshots of a subset of the Web of Data, to study the dynamics of linked data. They started crawling snapshots since May 2012, which are freely available to

[8] `https://www.w3.org/TR/vocab-adms/`, last accessed: November 28, 2019
[9] `https://sparontologies.github.io/cito/current/cito.html`, last accessed: November 28, 2019
[10] `http://www.w3.org/TR/vocab-data-cube/`, last accessed: November 28, 2019
[11] `https://www.w3.org/TR/vocab-dcat/`, last accessed: November 28, 2019
[12] `http://lov.okfn.org/dataset/lov/vocabs/emp`, last accessed: November 28, 2019
[13] `http://data.ign.fr/def/geometrie/20160628.htm`, last accessed: November 28, 2019
[14] `http://www.geonames.org/ontology/documentation.html`, last accessed: November 28, 2019
[15] `http://www.geonames.org/ontology/documentation.html`, last accessed: November 28, 20199
[16] `http://www.openannotation.org/spec/core/`, last accessed: November 28, 2019
[17] `https://www.w3.org/TR/vocab-org/`, last accessed: November 28, 2019
[18] `https://www.w3.org/TR/prov-o/`, last accessed: November 28, 2019
[19] `http://lov.okfn.org/vocommons/voaf/v2.3/`, last accessed: November 28, 2019
[20] `http://rdf-vocabulary.ddialliance.org/xkos.html`, last accessed: November 28, 2019

download[21]. For this chapter of the thesis, we parsed 242 snapshots (from May 2012 to March 2017).

The Billion Triple Challenge (BTC): BTC is part of the Semantic Web Challenge[22], with the primary goal of crawling a dataset from the LOD cloud. In this analysis, we used all available BTC datasets, which were crawled in the years 2009, 2010, 2011, 2012, and 2014, to analyze the use of vocabulary terms and the adoption of changes of the extracted vocabularies.

Wikidata: Wikidata is a knowledge base to collaboratively store and edit structured data. It is a storage for the structured data of Wikipedia, Wikivoyage, and others. Wikidata is a free and multilingual repository[23]. In order to analyze the Wikidata vocabulary, we first extracted the terms introduced by this vocabulary. Using the RDF exports from Wikidata page[24], we parsed the types and properties from the RDF dump files that were generated using the Wikidata toolkit[25]. We assumed that the first snapshot of those files was the first version of the Wikidata vocabulary, and based on this assumption, we parsed the next dump files to extract the changes to the first version, and so on. Overall, there were 25 RDF dump files (from April 2014 to August 2016). Using those files, we extracted the terms that were added or deprecated. Subsequently, we parse the Wikidata dump files to extract the use of its terms to analyze the adoption of changed types and properties of the Wikidata vocabulary.

[21] `http://km.aifb.kit.edu/projects/dyldo/`
[22] `http://challenge.semanticweb.org/`, last accessed: November 28, 2019
[23] `https://www.wikidata.org/`, last accessed: November 28, 2019
[24] `http://tools.wmflabs.org/wikidata-exports/rdf/exports.html`, last accessed: November 28, 2019
[25] `https://github.com/Wikidata/Wikidata-Toolkit`, last accessed: November 28, 2019

4.3 Results

In this section, we present the results of analyzing the use of vocabulary terms and adoption (Section 2.1) of these changes. Section 4.3.1 presents the results of the use of terms and the adoption of changes in the LOD cloud. Section 4.3.2 presents the findings regarding the use of the Wikidata vocabulary terms and the adoption of newly created types and properties in the Wikidata dataset [2].

4.3.1 Use and Adoption of Vocabulary Terms in DyLDO and BTC

This section presents the results of parsing the DyLDO and BTC datasets to extract information regarding the use of the 13 vocabularies (see Section 4.1). Furthermore, we present the findings of adopting the changes in vocabulary types and properties in these datasets.

Change and Use of LOD Vocabularies

To analyze the adoption time of the changes of LOD vocabularies, first, we need to understand the changes and use of vocabulary terms, focusing on two types of changes: creation and deprecation. We used LOV to download all the versions of the 13 vocabularies. Overall, we observed 35 % of newly created terms and 11 % deprecated ones. Furthermore, 85 % of the vocabularies had an increased number of types and properties, as shown in Figure 4.1. Two exceptions were *adms* and *cito*: the number of terms decreased for the former, while the latter vastly dropped the total number of types.

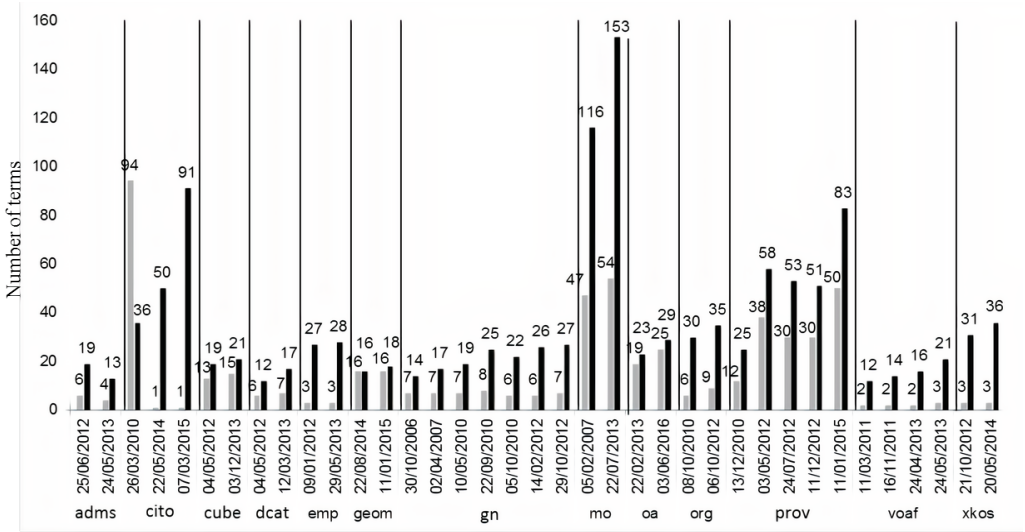


Figure 4.1: Changes in the vocabularies over time. The gray bar represents the total number of types and the black bar represents the properties for each of the selected vocabulary over their versions. The x-axis represents the versions of each vocabulary, and the y-axis shows the total number of types and properties for each version.

In the DyLDO, most vocabularies are used steadily. Thus, we show in Figure 4.2 the vocabularies with different changes in use. Notably, *mo* shows increasing and declining intervals, *prov* is increasing in popularity despite some slight negative picks, while *adms* had a significant drop in 2015 after an initial increase in use, although it slightly increased from 2015 to 2017. Furthermore, *cube* had a peak towards the end of 2015 and then returned to its initial use rate, while *emp* seems to have been unused since 2015.

The data publishers still use a great majority of the deprecated terms (87 %). We found that *geonames.org* is the PLD most frequently using deprecated terms in the BTC and DyLDO datasets. For instance, Figure 4.3 shows the use of the `gn:Country` type in DyLDO, which was deprecated in September 2010. Despite various fluctuations, its use increased until August 2015, then declined and increased again to reach a peak in August 2016.

Publishing new versions of vocabularies may influence the use of terms.

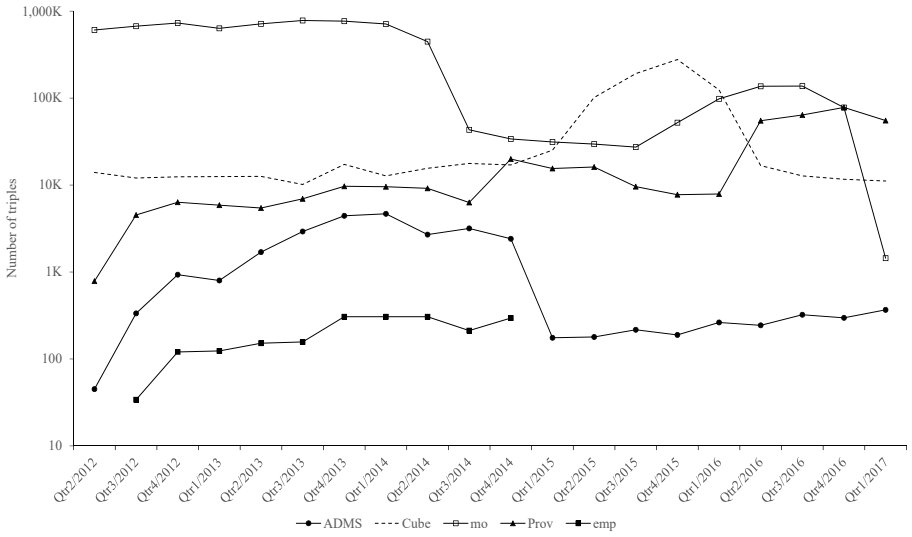


Figure 4.2: The mean number of triples that use terms of the *adms*, *cube*, *mo*, *prov*, and *emp* vocabularies in the DyLDO dataset (figures aggregated over quarters).

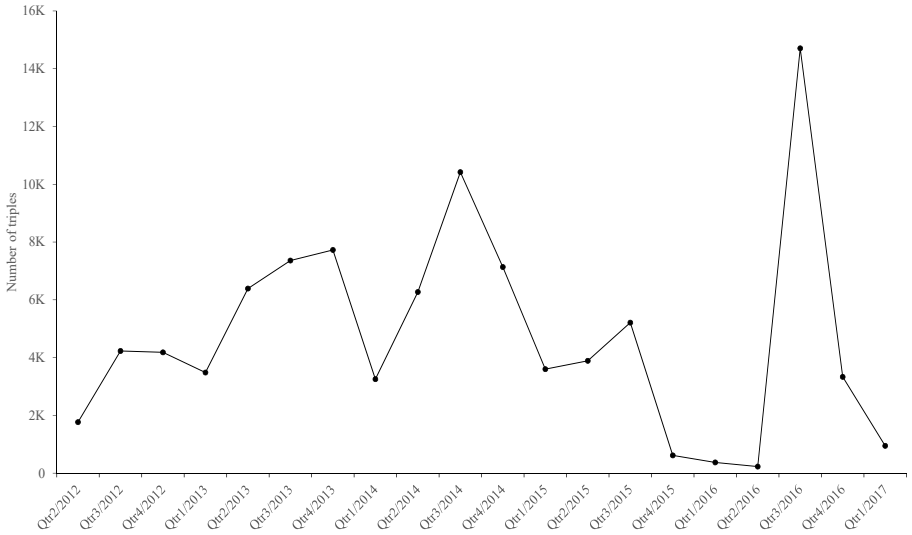


Figure 4.3: The use of the `gn:Country` type in the DyLDO dataset (figure aggregated over quarters).

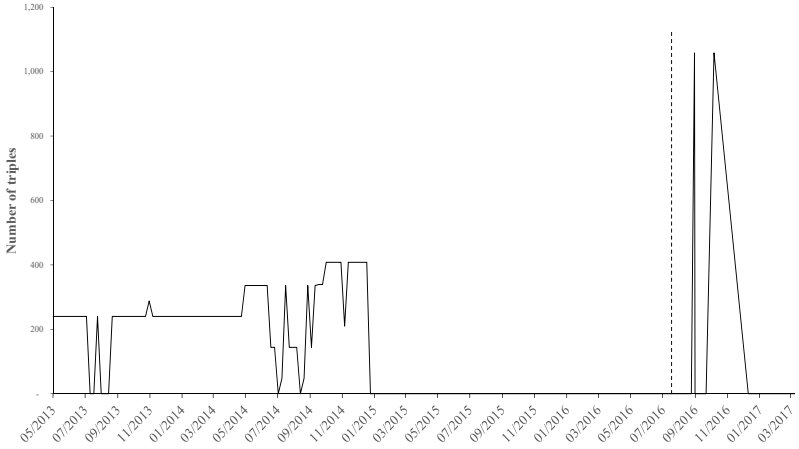


Figure 4.4: Amount of triples that use the *oa* vocabulary in the DyLDO dataset. The vertical dashed line represents the time of publishing the new version of the *oa* vocabulary.

Figure 4.4 illustrates the use of *oa* in the DyLDO dataset. The *oa* vocabulary published its first version in February 2013. When a new version was published in June 2016, the number of triples using *oa* increased. Although the number of triples reaches a peak point (about 1K) that may be considered small, we can still notice an influence of these changes on the number of triples using the *oa* vocabulary.

Furthermore, we can state that not all terms are used. For example, the percentage of the used terms for half of the vocabularies is less than 50 % of terms in the BTC dataset (in total, 50 % of all terms were not used). While in DyLDO, the percentage of unused terms was 23 %, and only one (*cito*) had a percentage of unused terms equaling 60 %. In comparison, the remaining vocabularies have a percentage of less than 40 (Table 4.2).

Table 4.2: The percentage of unused terms in the Billion Triples Challenge (BTC) and DyLDO datasets. The *Total terms* column represents the total number of terms the vocabulary created during its lifespan.

Vocabulary	Total terms	BTC	DyLDO
adms	31	68 %	3 %
cito	220	72 %	60 %
cube	37	35 %	0 %
dcat	23	48 %	9 %
emp	31	87 %	6 %
geom	34	100 %	3 %
gn	43	26 %	9 %
mo	208	36 %	2 %
oa	63	83 %	35 %
org	44	20 %	11 %
prov	143	22 %	24 %
voaf	24	33 %	8 %
xkos	35	63 %	14 %

Adoption of LOD Vocabulary Changes

The majority of the newly created terms were adopted in less than ten days, as shown in Table 4.3. Only the adoption of *geom* and *gn* took a long time. The triples column represents the total number of triples in DyLDO containing the adopted terms, while μ and σ are the average number of days before adoption and the standard deviation, respectively. Notably, the 21 new terms of the *oa* vocabulary and only one *xkos* new term were never adopted.

After being adopted, the use of 50 % of the newly created terms decreased during the considered period, 47 % showed a steady use, while 3 % further increased. For example, during its evolution, the *voaf* vocabulary created ten new terms. All but one of these terms saw a decline in the use, starting from the fourth quarter of 2014. Figure 4.5 shows only six terms. The remaining new terms were exploited in much fewer triples (less than 10 triples); thus, we did not include them in the figure. In general, a similar trend holds for all the vocabularies.

Table 4.3: The adoption of newly created types and properties for each of the vocabularies.

Vocabulary	New terms	Adopted terms	Triples	μ	σ
adms	6	100 %	31K	7	0
cito	80	100 %	281K	7	0
cube	5	100 %	15K	7	0
dcat	5	100 %	104K	8.4	3.13
emp	1	100 %	4K	7	0
geom	2	100 %	16K	420	0
gn	21	100 %	160M	127.76	255.33
mo	44	100 %	45M	8.75	9.68
oa	21	0 %	-	-	-
org	8	100 %	173K	7	0
prov	106	85 %	121M	30.15	37.49
voaf	10	100 %	75K	43.33	68.58
xkos	1	0 %	-	-	-

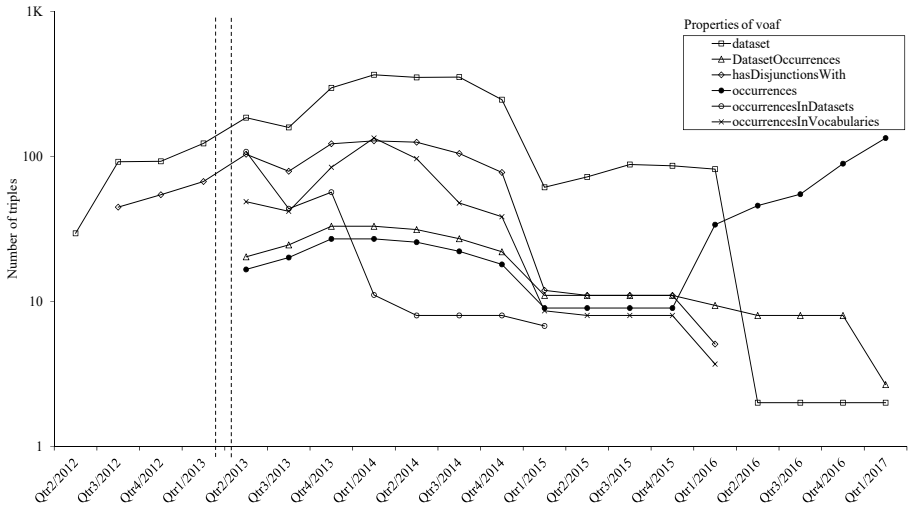


Figure 4.5: The amount of triples in which a *voaf*'s newly created type or property occurs per quarters of DyLDO snapshots. The vertical dashed lines represent the publishing time of new versions of the vocabulary. Please note that two versions of *voaf* have been published before the first snapshot of DyLDO (i. e. the properties `dataset` and `hasDisjunctionsWith` are newly created in versions released before the second quarter of 2012).

4.3.2 Use and Adoption of Vocabulary Terms of Wikidata

For analyzing the use and adoption of the Wikidata vocabulary, we parsed the types and properties from the 25 RDF dump files, for the period from April 2014 to August 2016. We extracted the added and deprecated terms of the Wikidata vocabulary. Figure 4.6 presents the total number of types and properties in each Wikidata snapshot, which constantly grew and reached 11 types and 27 properties in August 2017. Ontology engineers added three types and nine properties during the analyzed period. Notably, there are no terms deprecated during the ontology evolution. The new types are `DeprecatedRank`, `PreferredRank`, and `NormalRank`, while the new properties are `propertyTypeMonolingualText`, `propertyTypeProperty`, `propertyQualifierLinkage`, `propertyReferenceLinkage`, `rank`, `propertyStatementLinkage`, `propertySimpleClaim`, `quantityUnit`, and `propertyValueLinkage`.

Figure 4.7 illustrates the use of newly created terms in Wikidata. Only five out of 12 terms were adopted. `NormalRank` and `rank` are used much more than the other terms. Furthermore, the actually adopted terms among all the newly created ones were adopted directly after their creation date (i. e., on the same day). One possible reason is that Wikidata is a more controlled and centralized environment than the distributed LOD cloud, as discussed in Section 4.4.2.

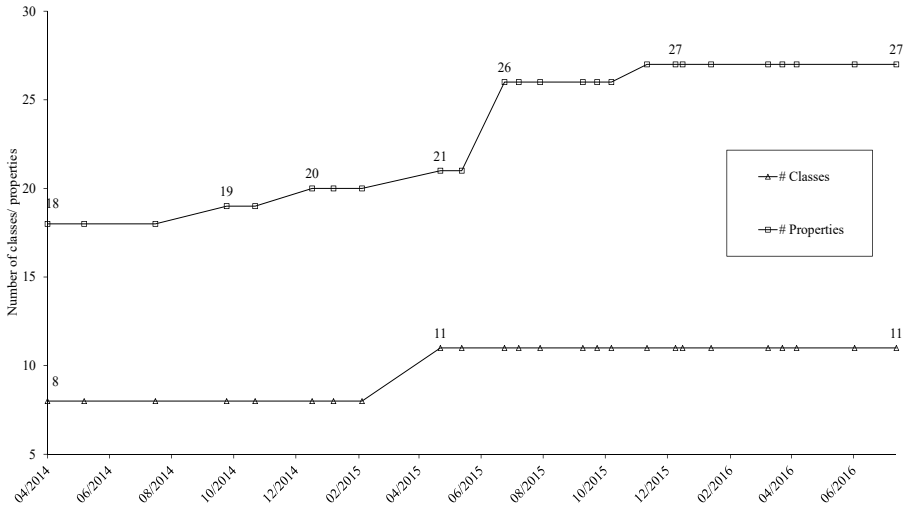


Figure 4.6: Total number of terms of the Wikidata vocabulary per RDF dump file.

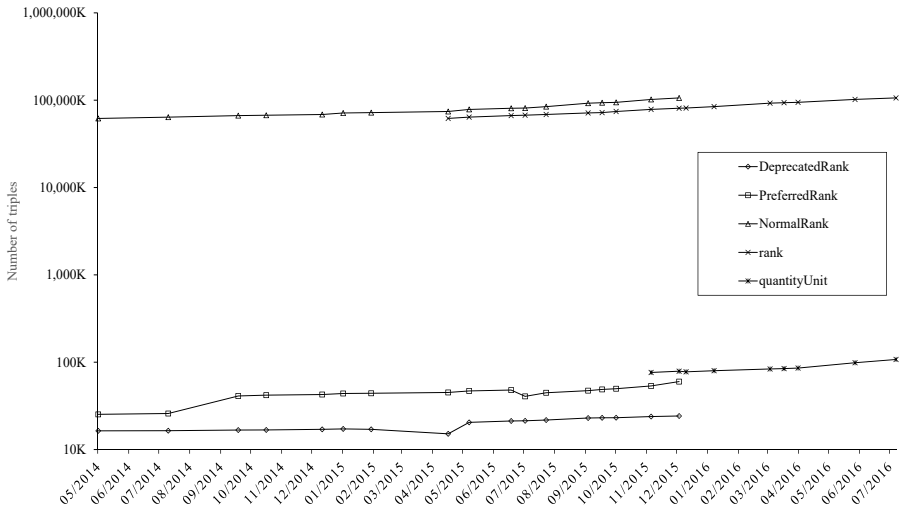


Figure 4.7: The number of triples that adopted the newly created types and properties of the Wikidata vocabulary.

4.4 Discussion

We found that not all vocabulary changes are reflected in the published data. Additionally, we found that data publishers still exploit most of the deprecated types and properties. The Wikidata vocabulary has no deprecations, and there is a huge difference in the number of triples in which the types and properties of Wikidata occur. In Section 4.4.1, we discuss in detail the results related to use and adoption of vocabulary terms in the LOD cloud. In Section 4.4.2, we discuss the results of changes and adoption of the Wikidata vocabulary terms.

4.4.1 Use and Adoption of Vocabulary Terms in DyLDO and BTC

Change and Use of LOD Vocabularies

The total number of terms increased in most of the vocabularies. This suggests that more knowledge is represented in the LOD cloud, i. e., new terms are required. One exception is *cito*, which consisted of 94 types and 36 properties when initially published. The second version counted only one type and 50 properties. Specifically, all the 94 types were replaced with the new type `CitationAct`, and most of the 36 properties of the first version were substituted. The third version provided 91 properties, although 18 of the new properties were reintroduced from the first version (deprecated in the second version). In fact, almost a new ontology was built. This is particularly important since *cito* has grown in popularity (BTC 2014 contained over 300,000 triples compared to the 40,000 in the BTC 2011) [2].

In general, our analysis shows that the number of changed types and properties is small. This is in line with existing studies [1, 26, 52]. However, these changes may have a large impact on the published data. For example, the new version of the *oa* vocabulary caused a significant increase of its use: the triples containing its terms almost triplicate (from roughly 400 to over 1,100). In general, the impact of the changes on the use of vocabulary terms, either in an increasing or decreasing way (six and five out of 13 vocabularies, respectively), although at different times. For *dcat*, there was a delay in use by three years.

Although some terms are deprecated, 87 % of them were still used. This is in line with Meusel et al. [52]. *geonames.org* is the PLD with the highest number of deprecated terms. For example, in the BTC 2011 dataset, *geonames.org* used six deprecated terms in about 522,000 triples. That number declined to three terms and roughly 181,000 triples in BTC 2012, but increased again to 49 terms in BTC 2014 (5,500 thousand triples). It seems that data publishers did not update their data models. A possible reason for this is that they are not aware of changes in the vocabularies. Thus, as previously discussed in Chapter 3, they could benefit from tools notifying these changes [2].

To provide information about the status of a term, the *Vocabulary Status Ontology*[26] can be used. This ontology consists of three properties: `vs:term_status`, `vs:moreinfo`, and `vs:userdocs`. Unfortunately, it is not widely used, only seven out of the 134 vocabularies that were investigated in this study relied on it.

[26] `https://www.w3.org/2003/06/sw-vocab-status/note.html`, last accessed: November 28, 2019

Adoption of LOD Vocabulary Changes

Most of the newly coined terms were adopted rather quickly (in less than one week). Surprisingly, we even found some terms being adopted even before their official publishing date [2]. We believe that some of the new versions of vocabularies are already online, and thus can be used before their official announcement. Another possibility is that the data publishers are the same as vocabulary engineers. In some cases, it may take time to publish the new version of the vocabulary.

Although most of the terms were quickly adopted, some of the newly created terms, such as terms of *gn*, took more than 120 days, on average, to be adopted. However, this average does not reflect the actual adoption behavior of a vocabulary. The new version of *gn* provides 21 new terms, 17 terms of which were adopted within seven days, while the remaining four terms were adopted after 600 days. Therefore, the average result was affected by those few terms with the long adoption time.

Another interesting point is that some newly created terms are never adopted. For example, ontology engineers published a new version of the *oa* vocabulary in June 2016, with 21 new types and properties. None of those terms have yet been adopted (until April 2017, the last DyLDO snapshot considered in this work), while the first version of *oa* was published in February 2013 with 42 terms and all but one were adopted in less than three months. However, the reasons why those terms are unused likely depends on the specific application scenario. For instance, not all terms need to be currently used, and some could be designed for future applications. Furthermore, although some terms are not used in the LOD cloud, they may be exploited in other forms, such as they may be used to define the properties' types [2]. We do not claim that every term has to be adopted. We

also believe that raising awareness for data publishers about the existence of new terms in an ontology in use may further stimulate the use of the terms.

4.4.2 Use and Adoption of Vocabulary Terms of Wikidata

We found that the Wikidata vocabulary showed no deprecated types or properties, although some were never adopted during the investigated time-frame (e. g., the `Article` type) [2]. The Wikidata vocabulary, like most of the LOD vocabularies, counted a small number of additions (three types and nine properties) and no deprecation.

Three types (`DeprecatedRank`, `NormalRank`, and `PreferredRank`) suddenly disappeared from Wikidata statements after the snapshot in December 2015, after about eight months (they were created in May 2015). We think that the ontology engineers of Wikidata retain the unused terms even if the data publishers stop using them. There is a huge difference in the number of triples that use a term. For instance, the `NormalRank` and `Statement` types have been used in about 106 and 81 million triples, respectively. The other types (except `Item`) have been used in less than 2.4 million triples. For `Item`, it was used in around 19 million triples on average. The same observation can be made for the properties: all but `rank` appeared in less than 2.7 million triples, while `rank` accounted for approximately 62 million triples, when introduced in May 2015, then reached about 106 million triples in August 2016. The wide exploitation of these terms suggests a pressing necessity for adding them to the vocabulary.

Surprisingly, the majority of new terms (two types and nine properties of in total 12 terms) were not adopted in any statements of the Wikidata. However, a more

in-depth analysis showed that the un-adopted terms are used to define properties and their types, except the `Article` type, which needs further investigation. The five newly created terms that are used in Wikidata were directly adoption after their creation date. This was more expected in Wikidata than for the LOD cloud, because the former is a controlled and centralized environment versus the distributed LOD cloud. Furthermore, the data publishers of Wikidata are the same as vocabulary engineers.

4.5 Summary

In this chapter, we analyzed the use and adoption of vocabulary terms for modeling data (RQ2, introduced in Section 1.2). We first quantified the amount and frequency of changes in a set of vocabularies. Subsequently, we investigated to which extend and when the changes were adopted in the data. The conducted methodology resulted in 13 vocabularies with more than one version and at least one version published after May 2012, which were investigated using some well-known datasets. We conducted our experiments on three large-scale datasets for which time-stamped information is available, namely the BTC, DyLDO, and Wikidata.

The data publishers still reuse 87% of the examined deprecated and deleted terms from the 13 vocabularies, based on the DyLDO and BTC datasets. Thus, data publishers may not be aware of changes in the vocabularies, and thus exploit the reuse. In less than a week, most of the newly coined types and properties were adopted by data publishers. There were some exceptions, but this is a general observation. Although there is generally fast adoption of newly created terms,

some new terms are never adopted. The reasons why those terms are not used are not explicitly known; they may be used in a specific application scenario.

There exist some vocabularies that provide information about the status of the types and properties of the vocabularies, such as the SemWeb Vocabulary Status ontology (vs). We recommend that ontology engineers use these status vocabularies in order to provide the current status of the vocabulary terms.

For the Wikidata vocabulary, we found that there were no deprecations or deletions of its types or properties. Wikidata, like most of the LOD vocabularies, showed a small number of additions in its terms. The extensive exploitation of the newly coined terms of the Wikidata vocabulary suggests a pressing necessity for adding them to the vocabulary, due to the enormous number of triples in which the types and properties occur.

Chapter 5

Analysis of the Network of Linked Vocabularies (NeLO)

It is common practice to reuse existing terms, i. e., types and properties, defined in the vocabularies for modeling one's data[1]. The goal is preventing the proliferation of terms and reducing the range of choices when modeling data. Thus, ontology engineers should import some terms from other vocabularies if they fit their needs, instead of creating new ones [35]. This reuse of terms leads to a NeLO.

Changes in one or more vocabularies of the NeLO may lead to a problem regarding their dependencies. A dependency is a vocabulary that reuses one or more types or property from another vocabulary. Additionally, the changes influence the published data on the web as well. The impact of vocabulary changes in data has been discussed in Chapter 4.

In this chapter, we present our analysis of the vocabulary changes at the schema

[1] `http://linkeddata.org/`

level and their impact on the other vocabularies in the NeLO. We extend the methodology that we used to analyze the vocabulary changes and discussed in Chapter 3. We use again the LOV dataset[2] and analyze 994 vocabularies and their changes over 17 years. The vocabularies are considered as part of the network if they import or export at least one term from some other vocabularies. A broad range of network-analysis metrics were employed for the extracted network. The metrics were continuously applied during the evolution of the NeLO to find out how the important nodes have changed over time.

The remainder of this chapter is structured as follows: Section 5.1 includes the methodology to analyze the evolution of the NeLO. The results of our analysis on the vocabularies of the LOV dataset are listed in Section 5.2. In Section 5.3, we discuss the findings of analyzing the evolution of NeLO and conclude.

5.1 Network Analysis Methodology

5.1.1 Procedure

To answer the research questions related to the evolution of the NeLO, we extended the methodology applied in Section 3.1.2 to analyze the changes of vocabularies by applying multiple network-analysis metrics to analyze the NeLO, and repeating these metrics to analyze its evolution (steps 2 and 4 in the methodology below). We used the same 636 vocabularies and their types and properties from LOV as in Chapter 3. Note that while analyzing the 636 vocabularies, some additional vocabularies not contained in the LOV were found.

[2] `http://lov.okfn.org/dataset/lov/`, last accessed: November 28, 2019

Thus, we considered a total of 994 vocabularies. The extended methodology consists of the following steps:

1. We extracted all types and properties from all the available versions of vocabularies (from June 2001 to June 2018), listed in the LOV dataset. The terms extracted were classified into two categories: the *own terms*, created by the ontology engineers of the considered vocabulary, and *imported terms* are reused from other vocabularies.

2. We employed multiple network-analysis metrics to study the NeLO and its evolution. The network-analysis metrics are described in more detail in Section 5.1.2.

3. We checked whether the *imported terms* were the most recent ones, i. e., whether they appear in the latest published version of the source vocabulary and are actually reused in the target vocabulary; otherwise, older versions are considered instead.

4. We repeated the first and second steps on the evolving NeLO to analyze the change in it over time. The process of selecting the snapshots was done every June, starting from 2017 back to the year of the first available vocabulary, which was in 2001.

The details of the first and third steps of the methodology can be found in Section 3.1.2. For the second step of the methodology, we used the Open Graph Viz Platform (Gephi)[3] version 0.9.2 to visualize and analyze the NeLO. Gephi is an open-source tool to explore and analyze graphs,it provides the most common metrics for social network analysis and scale-free networks, such as Centrality,

[3] `https://gephi.org/`, last accessed: November 28, 2019

Closeness, HITS, Clustering Coefficient, PageRank, and others. The first two steps were repeated to analyze the evolution of NeLO over time. This process represents the fourth step of the methodology.

5.1.2 Metrics

In this section, we summarize the main metrics exploited for studying the NeLO. Network-analysis metrics, such as PageRank and HITS, can help to identify nodes which can be problematic because they have many dependencies, or their changes may affect many other nodes because they are widely reused. We exploited these measures in addition to the degree since they consider indirect dependencies (indirect links in NeLO).

Graph Density. The density of a graph is the actual number of edges in a graph over the maximum number of edges possible among all the nodes [14]. For a directed graph, the graph density (D) is defined by Equation 5.1.

$$D = \frac{|\,E\,|}{|\,V\,|\,(|\,V\,|\,-1)} \tag{5.1}$$

Degree, in-degree, out-degree. The degree of a node is the number of edges, both incoming and outgoing, connecting it to other nodes; the in-degree of a node is the number of its incoming edges, the out-degree of a node is the number of its outgoing edges [6]. In our model of the NeLO, the in-degree corresponds to the number of vocabularies from which a given ontology imports at least one term. Analogously, the out-degree represents the number of vocabularies to which a given ontology exports at least one term.

Average degree. In a directed graph, the graph average degree is the total number of edges divided by the number of nodes [20]. The average degree $\langle k \rangle$ for a directed graph is defined by Equation 5.2.

$$\langle k \rangle = \frac{|\,E\,|}{|\,V\,|} \tag{5.2}$$

Network diameter. The diameter of a network is the longest path from the set of all shortest paths between all the nodes in a graph [7]. To calculate the network diameter, first, we calculated the shortest paths for each pair of nodes in the graph, and then we found the longest path from the resulting scores.

PageRank. This metric calculates the ranks for every node in a graph to assess its importance [60]. The PageRank score for a node v is defined by Equation 5.3, where $v, v' \in V$ and v' is connected to v. B_v is the set of all nodes that have edges pointing to v, $d_{out}(v')$ is the outgoing degree of the node v', and $PR(v')$ is the PageRank score for the node v'.

$$PR(v) = \sum_{v' \in B_v} \frac{PR(v')}{d_{out}(v')} \tag{5.3}$$

Hypertext Induced Topic Selection (HITS). This metric analyzes nodes by their incoming and outgoing degrees [45]. Nodes that point to other nodes are called *hubs*, and the nodes that are pointed from other nodes are called *authorities*. Hubs and authorities are given scores based on their incoming and outgoing degrees. The hub and authority scores of a node v, $hub(v)$ and $auth(v)$ are respectively defined by Equation 5.4 and

Equation 5.5, where i is a node that has an edge pointing to the node v.

$$hub(v) = \sum_{i=1}^{n} auth(i) \tag{5.4}$$

$$auth(v) = \sum_{i=1}^{n} hub(i) \tag{5.5}$$

Degree Centrality. This metric identifies the most important nodes in a graph [21]. The centrality algorithm specifies that the node with the highest degree is the most important one. The centrality of a graph G, i. e., $C(G)$, is defined by Equation 5.6, where v_c is the node with the highest degree of centrality.

$$C(G) = \frac{\sum_{i=1}^{|V|} C(v_c) - C(v_i)}{|V|^2 - 3|V| + 2} \tag{5.6}$$

5.2 Results

In Section 5.2.1, we present the results related to the state of the NeLO as of June 2018, and related to the evolution are listed in Section 5.2.2 [3].

5.2.1 Network of Linked Vocabularies in 2018

Figure 5.1 shows the current state of the NeLO after extracting all import relations between the latest versions of the vocabularies until June 2018. One can see three main circles in the network. These circles are formed depending on the number of exports to the other vocabularies. The central circle contains the vocabularies that have the most exports (more than 100 edges), which are represented by the

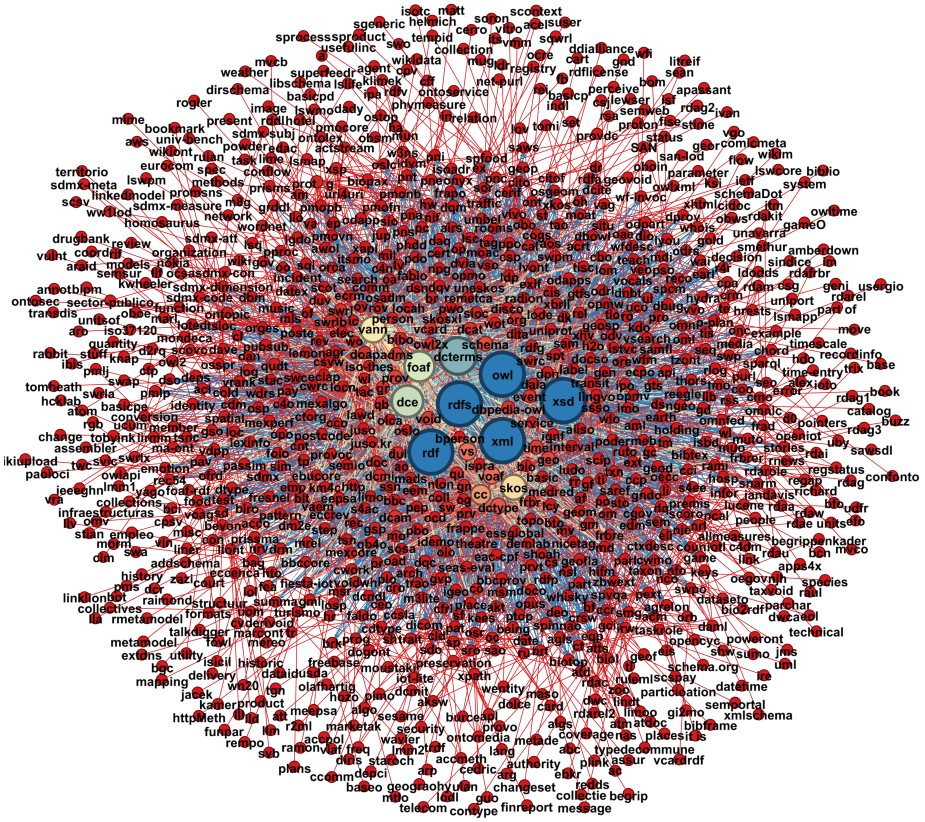


Figure 5.1: The Network of Linked Vocabularies as of June 2018.

larger node sizes. The middle circle (the denser area of smaller nodes) includes the vocabularies which have between five and 100 edges. The outer circle (the sparser external area) contains all the vocabularies that have been imported by less than five vocabularies.

Table 5.1 shows the basic NeLO's statistics for 2018. The table provides an overview of the state of NeLO by summarizing the relations between all the vocabularies, as of June 2018. Tables 5.2, 5.3, and 5.4 list the top-10 vocabularies that have the highest scores in Degree, HITS, and PageRank, respectively. We can see that the same vocabularies appear in most of the results of these metrics, but with some differences in their order. Furthermore, many meta-vocabularies, such as *owl*, *rdf*, *rdfs*, and *dce*, are in the top-10 for all the measures.

Table 5.1: Basic statistics of NeLO as of June 2018.

Measure	Value
Nodes	994
Edges	7046
Network diameter	12
Density	0.007
Average degree	7.089

Table 5.2: Top-10 vocabularies regarding degree, in-degree, and out-degree metrics in 2018, sorted by degree scores. The scores are calculated based on the import relationships.

Vocabulary	Degree	In-degree	Out-degree
owl	544	538	6
rdf	543	538	5
rdfs	543	538	5
xml	539	539	0
xsd	539	539	0
dcterms	435	425	10
dce	347	339	8
foaf	330	317	13
vann	255	244	11
skos	235	229	6

Table 5.3: Top-10 vocabularies for HITS scores in 2018, sorted by Authority.

Vocabulary	Authority	Hub
xml	0.368882	0.000000
xsd	0.368882	0.000000
rdf	0.368439	0.027594
rdfs	0.368439	0.027594
owl	0.368438	0.027628
dcterms	0.305421	0.037978
dce	0.242374	0.037727
foaf	0.234664	0.044112
vann	0.184754	0.045030
skos	0.171827	0.034529

Table 5.4: Top-10 vocabularies for PageRank scores in 2018.

Vocabulary	PageRank
xml	0.066215
xsd	0.066215
owl	0.057998
rdf	0.056593
rdfs	0.056593
dce	0.045954
dcterms	0.027649
skos	0.017678
foaf	0.013986
dcam	0.009152

Table 5.5: Top-10 vocabularies for degree, in-degree, and out-degree in 2018, sorted by degree. The scores are calculated based on the import relationships after excluding the meta-vocabularies.

Vocabulary	Degree	In-degree	Out-degree
dcterms	435	425	10
dce	347	339	8
foaf	330	317	13
vann	255	244	11
skos	235	229	6
cc	153	146	7
voaf	121	103	18
vs	116	108	8
dctype	82	74	8
schema.org	73	61	12

After excluding the meta-vocabularies from the previous results of the top-10 vocabularies, the scores of Degree, HITS, PageRank are shown in Tables 5.5, 5.6, and 5.7, respectively. Most of the vocabularies are the same as in the top-10 list of all these metrics, with some differences in their order. Furthermore, *dcterms*, *dce*, *foaf*, *skos*, and *vann* appear in the top of all three tables.

Figure 5.2 shows the number of vocabularies based on the average number of out-degree scores. We notice that the vocabularies that have nine versions have the highest out-degree scores (around 80). Furthermore, the second-highest

Table 5.6: Top-10 vocabularies for HITS scores in 2018, sorted by Authority. The scores are calculated after excluding the meta-vocabularies.

Vocabulary	Authority	Hub
dcterms	0.305421	0.037978
dce	0.242374	0.037727
foaf	0.234664	0.044112
vann	0.184754	0.045030
skos	0.171827	0.034529
cc	0.113386	0.034723
vs	0.081972	0.040256
voaf	0.080739	0.045920
dctype	0.058152	0.037727
schema.org	0.046659	0.040364

Table 5.7: Top-10 vocabularies for PageRank in 2018. The score are calculated after excluding the meta-vocabularies.

Vocabulary	PageRank
dce	0.045954
dcterms	0.027649
skos	0.017678
foaf	0.013986
dcam	0.009152
vann	0.009117
grddl	0.008740
dctype	0.005744
cc	0.005446
vs	0.005005

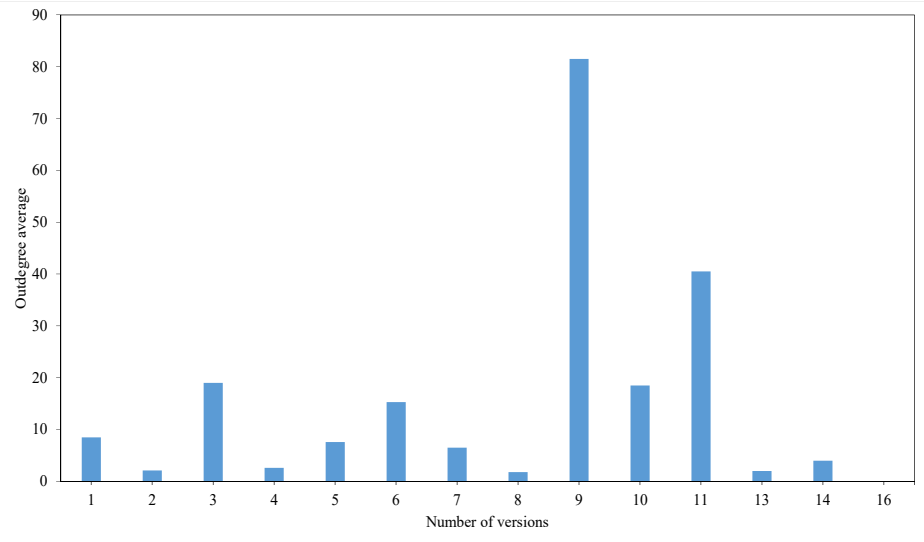


Figure 5.2: Distribution of vocabularies based on the number of versions and their out-degree scores.

category are the vocabularies that have 11 versions, which is an out-degree score of around 40.

Figure 5.3 represents the average number of exported terms, based on the out-degree scores. The numbers are calculated by dividing the number of exported terms by the out-degree score. These numbers show how many terms vocabularies export (on average) for other vocabularies. We can notice that the highest scores are for the vocabularies that have exported (on average) less than five terms to other vocabularies. Furthermore, Figures 5.4 and 5.5 show a histogram of the distribution of vocabularies, based on their in- and out-degree scores, respectively. Figure 5.4 shows that the highest number of out-degree scores is five, with around 700 vocabularies having this score, i. e., 700 vocabularies have been imported by another five vocabularies, on average. Furthermore, there are five vocabularies that have an out-degree of 600. These

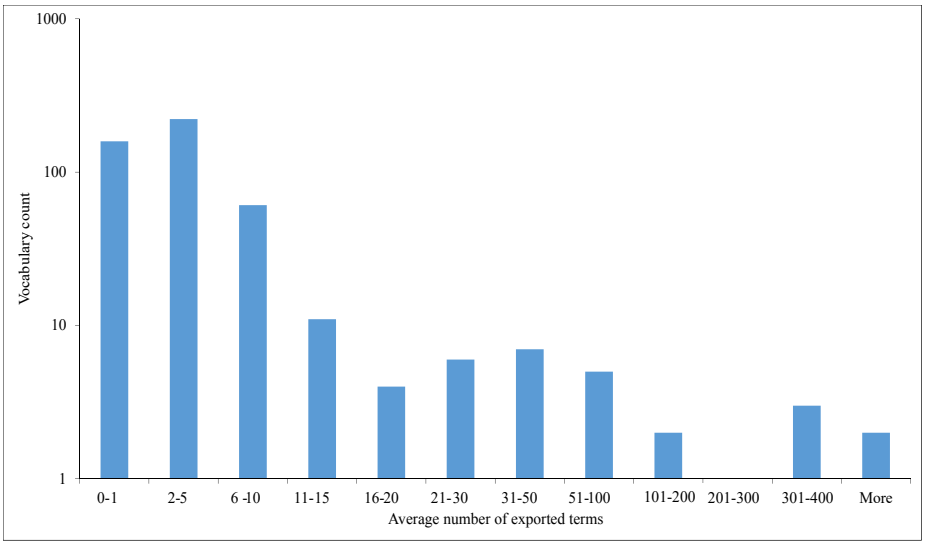


Figure 5.3: Average number of exported terms for each vocabulary in NeLO 2018. Note that the y-axis is log-scale.

vocabularies are meta-vocabularies. In Figure 5.5, around half of the vocabularies (450) have zero as in-degree scores.

5.2.2 Changes in the Network of Linked Vocabularies

Over time, the NeLO shown many changes in terms of the number of new vocabularies (nodes) and relations (edges) between them. Figures 5.6, 5.7, and 5.8 show 18 snapshots for the NeLO, from 2001 until 2006, from 2007 until 2012, and from 2013 until 2018, respectively. The figures illustrate the evolution of NeLO over time. We can notice that the network in 2002 (Figure 5.6a) started with 11 vocabularies and 30 edges in between. In 2017, Figure 5.8f shows that the network increased in size to a total of 958 vocabularies and 6,731 edges.

Figure 5.9 shows the total number of available types and properties in each NeLO

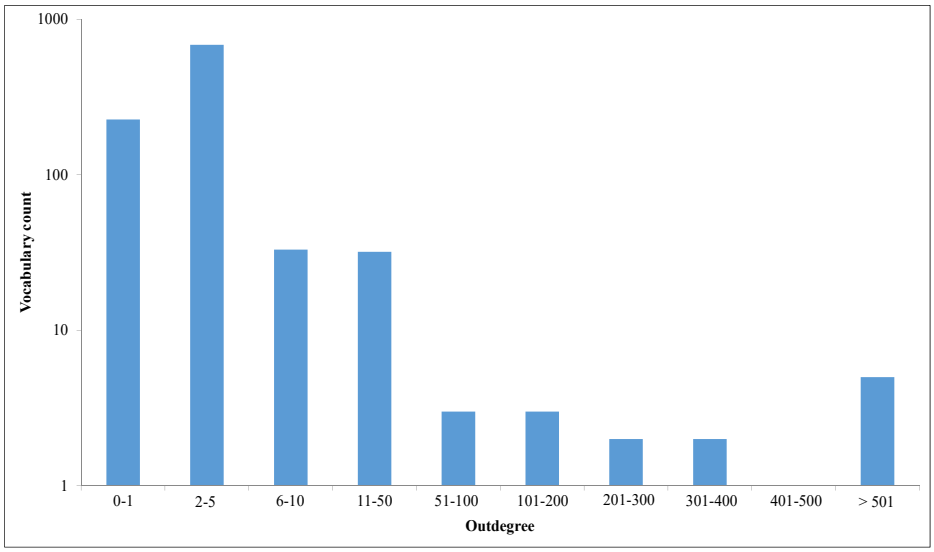


Figure 5.4: Distribution of out-degree scores for the vocabularies in NeLO 2018.

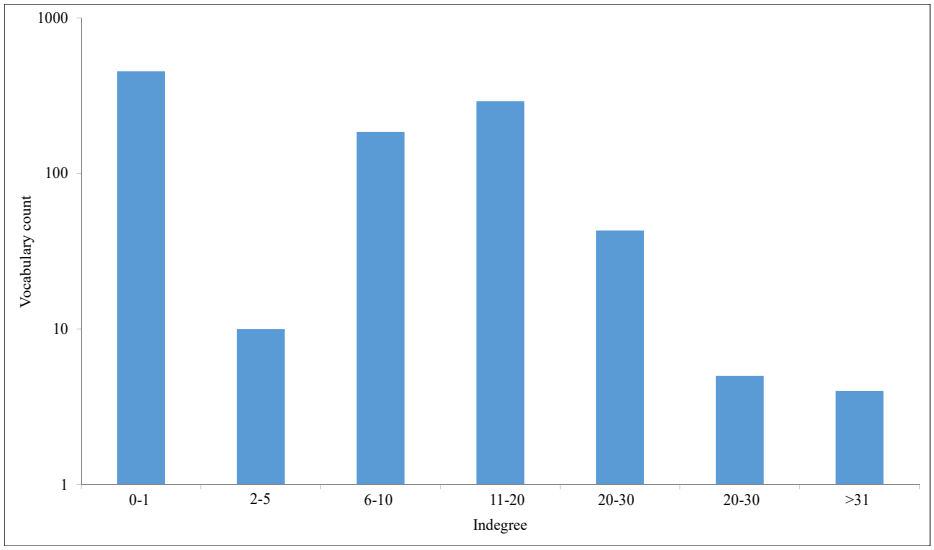
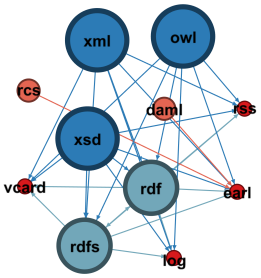
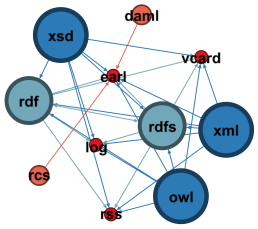


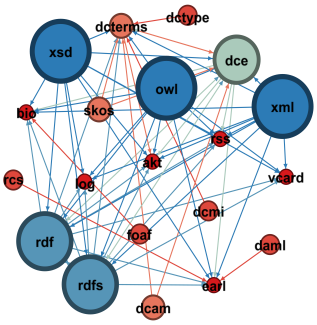
Figure 5.5: Distribution of in-degree scores for the vocabularies in NeLO 2018.



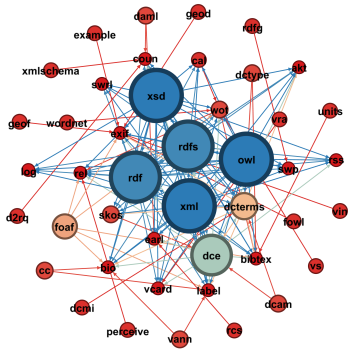
(a) NeLO as of 2001



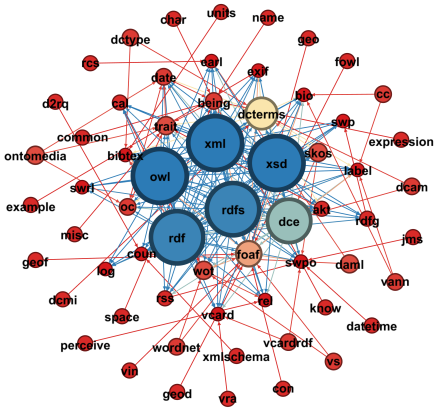
(b) NeLO as of 2002



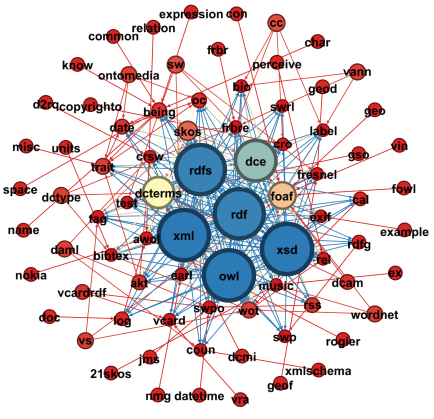
(c) NeLO as of 2003



(d) NeLO as of 2004

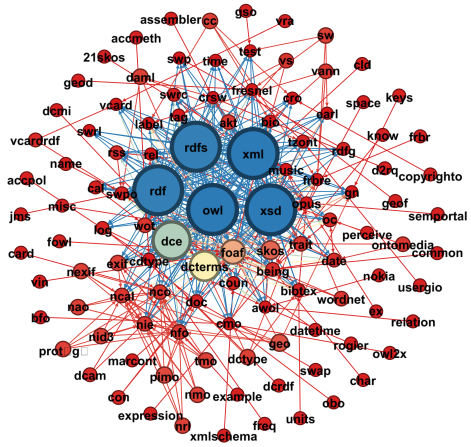


(e) NeLO as of 2005

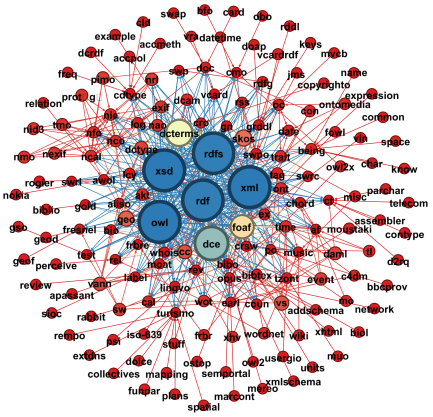


(f) NeLO as of 2006

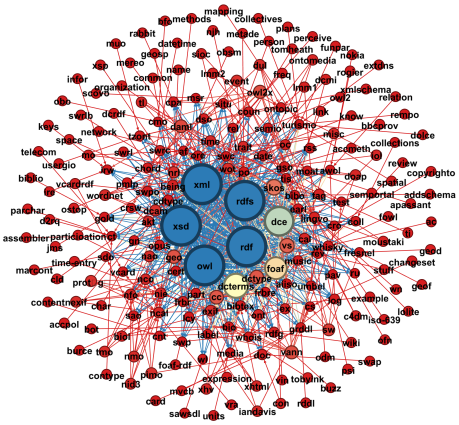
Figure 5.6: The evolution of the Network of Linked Vocabularies over time. The figures *a* to *f* represent six snapshots of NeLO from 2001 until 2006.



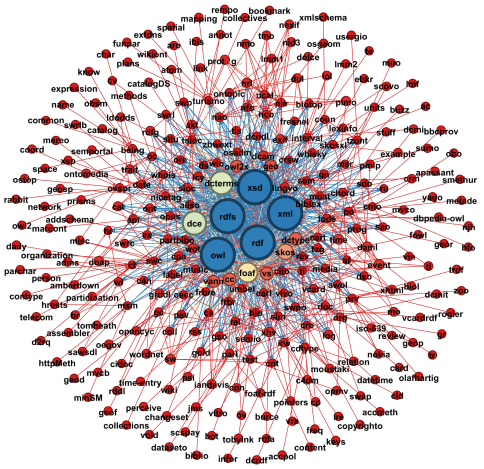
(a) NeLO as of 2007



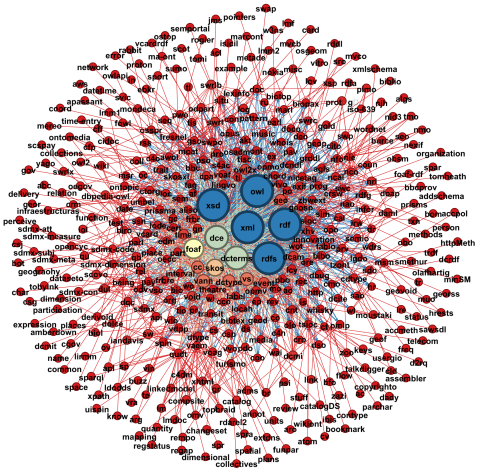
(b) NeLO as of 2008



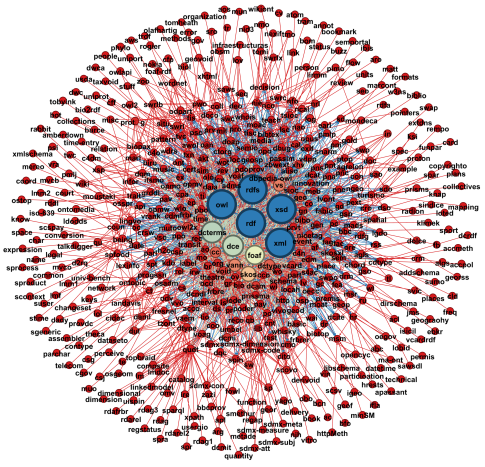
(c) NeLO as of 2009



(d) NeLO as of 2010

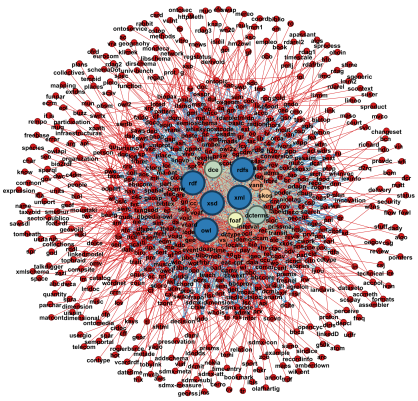


(e) NeLO as of 2011

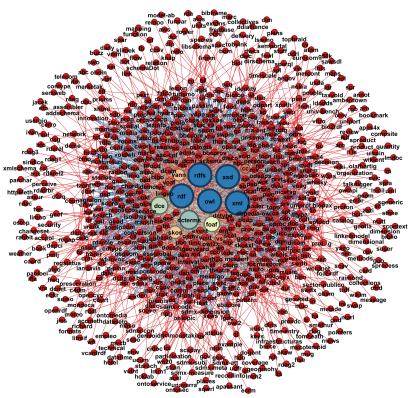


(f) NeLO as of 2012

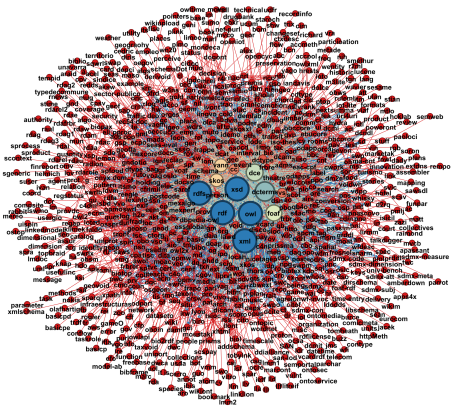
Figure 5.7: The evolution of the Network of Linked Vocabularies over time. The figures *a* to *f* represent six snapshots of NeLO from 2007 until 2012.



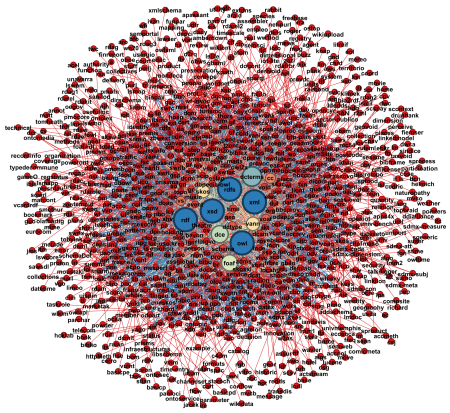
(a) NeLO as of 2013



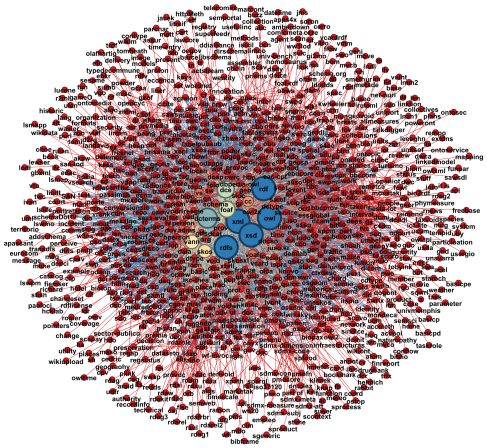
(b) NeLO as of 2014



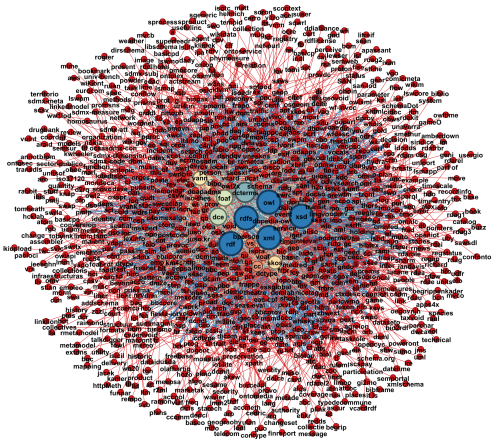
(c) NeLO as of 2015



(d) NeLO as of 2016



(e) NeLO as of 2017



(f) NeLO as of 2018

Figure 5.8: The evolution of the Network of Linked Vocabularies over time. The figures *a* to *f* represent six snapshots of NeLO from 2013 until 2018.

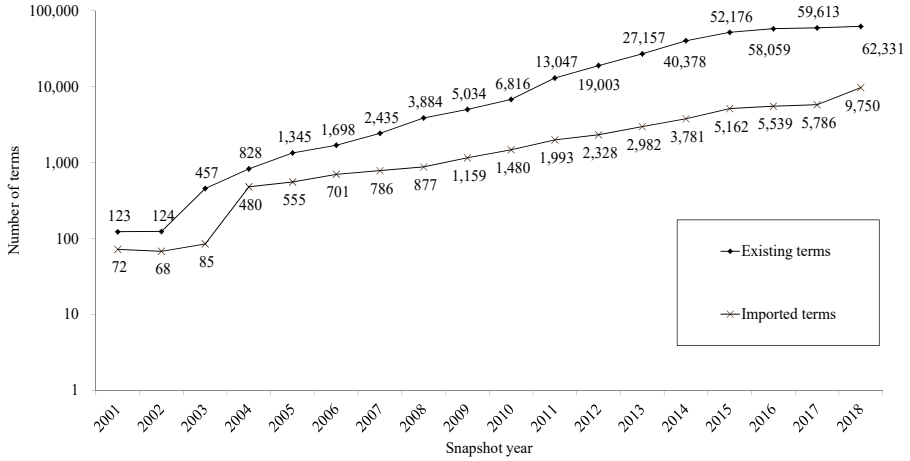


Figure 5.9: The total number of the existing terms in the NeLO vocabularies and the imported terms over time.

snapshot, and the total number of reused terms. The reuse percentage was at its peak with 10% and 11% in 2010 and 2011, respectively, while all other snapshots remain in the range between 5% and 7%.

Figure 5.10 depicts the total number of nodes and edges for each NeLO snapshot. It is worth noting that the number of nodes (vocabularies) and edges almost doubled in the 2003 and 2004 snapshots, compared to 2002 and 2003, respectively. Then, they continued to double every two years until 2013. Afterwards, the growth-rate decreased, and we can notice that since 2016 to June 2018, the number of new vocabularies that entered the network lowered (around 70 new vocabularies per year), while the number of new links was still slightly higher (about 600 per year).

Figure 5.11 presents an analysis of the NeLO over time that considers the density, network diameter, and average degree measures. We can notice that the average

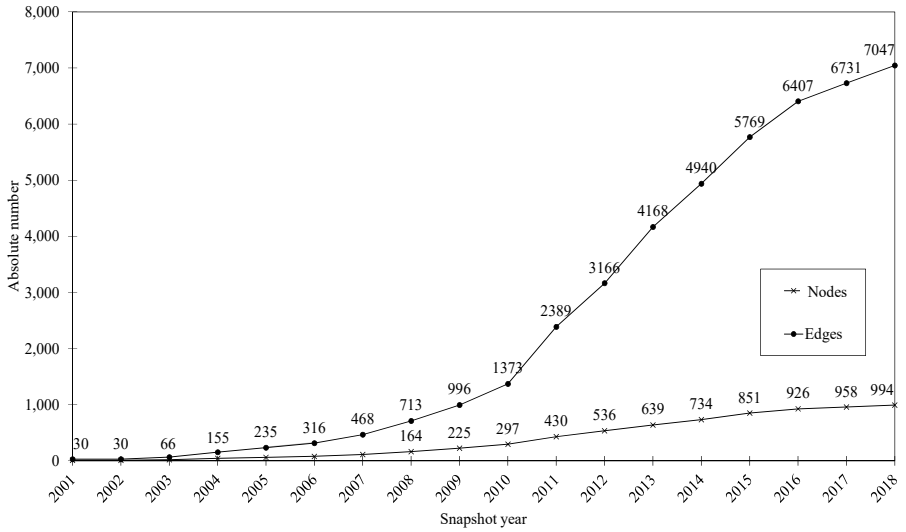


Figure 5.10: The evolution of NeLO. The figure shows the total number of nodes and edges for each NeLO snapshot until June 2018.

degree of the network has a slow but steady increase, while the density decreased over time. More specifically, in 2001, the network density was 0.273, and in 2018, it was 0.007. The network diameter sharply grew over the period considered. Specifically, first it quadrupled from 2002 to 2003, then there is another small peak from 2004 to 2005. From 2010 to 2015, we see the highest growth. Notably, the diameter of 2015 also represents also the maximum value in the whole period. Finally, in the last three years,the diameter has been almost constant.

Figure 5.12 presents the evolution of the most dominant vocabularies in terms of their out-degree. The out-degree corresponds to the total number of other ontologies that import at least one term from those vocabularies, i. e., the number of exports to different ontologies. The trend of the considered vocabularies is somewhat similar, although the absolute values differ. Specifically, the increase was low until 2010, and then comparably high until 2016. The vocabularies which import the most dominant vocabularies continue to grow after 2016, but in

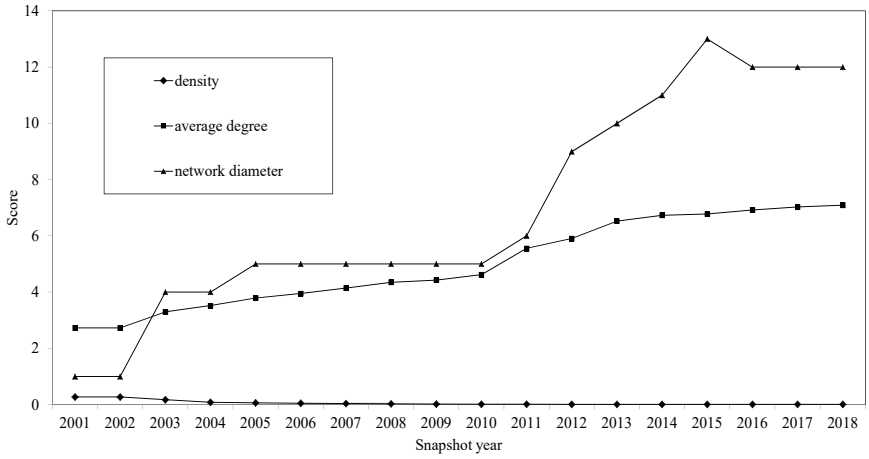


Figure 5.11: Change of NeLO in terms of density, average degree, and network diameter from 2001 to 2018.

a less pronounced way. Additionally, from 2003 until 2006, various vocabularies have a convergent out-degree score; the vocabularies *foaf*, *dce*, and *dcterms* are very close to each other until 2012. The vocabularies *dce* and *dcterms* were imported by the same number of vocabularies from 2010 to 2012. Afterwards, *dcterms* was imported by more vocabularies than *foaf* and *dce*. The *rdf*, *rdfs*, and *owl* vocabularies account for the highest growth in the whole period, with exactly the same pattern of being imported over the whole period. From 2007, they started being imported more often than the other three vocabularies considered. This gap continued to increase in the remaining period of time.

Figures 5.13a, 5.13b, and 5.14 illustrate the evolution of the in-degree, out-degree, and degree metrics for the top-five vocabularies, respectively, after excluding the meta-vocabularies. The selection of the top-five vocabularies is based on the last four NeLO snapshots (from 2015 to 2018). The in-degree was mostly constant up to a specific year, after which it increased. For example, in

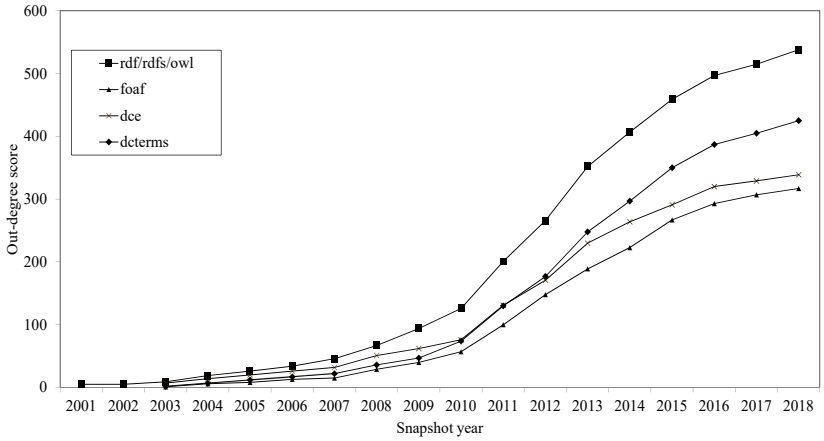


Figure 5.12: The change in the out-degree of the most dominant vocabularies over time. The *rdf*, *rdfs* and *owl* vocabularies are merged because they almost have the exact values.

2011 for *mo*, in 2015 for *interval*, and 2018 for *semio*, with some exceptions. In addition, we can notice that *qudt* decreased the number of imported vocabularies; in 2012, the number of imported vocabularies was 44. Subsequently, the number continuously decreased to 25 imported vocabularies in 2018. Furthermore, the *oa* vocabulary decreased the number of imported vocabularies from 23 in 2013 to only nine in 2016. Subsequently, this number increased again to 27 imported vocabularies in 2017. The *mo* vocabulary shows a constant number of imports from 2011 until the latest NeLO snapshot. It was introduced in 2007, and it does not import any term from the other vocabularies until 2011.

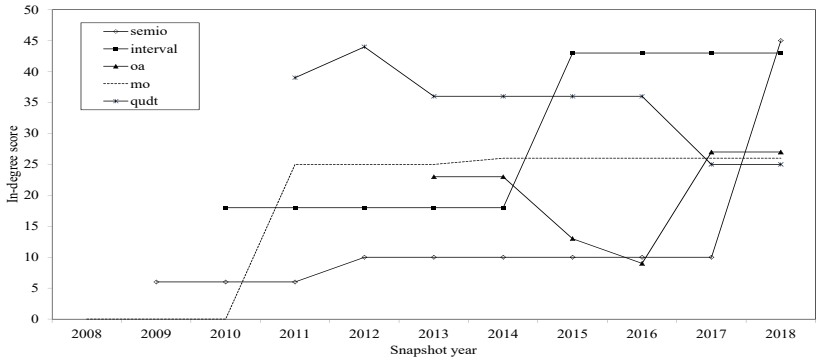
Figure 5.13b presents the out-degrees for the top-five vocabularies. From 2003 to 2007, all the vocabularies shown have similar out-degrees. From 2009, *skos* started to increase more than the others, and the same holds for *vann* from 2012. We can notice that *vann* and *skos* became widely popular, far more than the other vocabularies. Additionally, from 2015, *vann* exceeded *skos*, while before, *skos*

had the highest out-degree overall. However, the gap between their out-degrees is rather small. In 2014, *cc* achieved about the same out-degree as that of *vs*, and later on *cc* had a higher value than *vs*. The *voaf* vocabulary was introduced in 2011, and 2018, accounted for almost the same out-degree as *vs*.

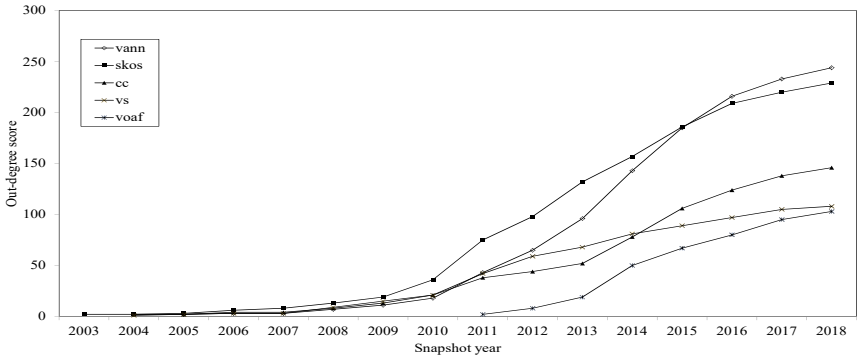
Figure 5.14 shows the degree scores for the top-five vocabularies, after excluding the meta-vocabularies. Notably, all the listed vocabularies' degree kept increasing, especially for *vann* and *skos*. As expected, the overall trend for each vocabulary is very similar to the out-degree, because the in-degree scores are much lower.

Figures 5.15 and 5.16 show the PageRank and HITS scores, respectively, of the top-five vocabularies selected for the degree analysis. In Figure 5.15, we can notice that all vocabularies have decreasing PageRank scores except *skos* and *vann*. The *skos* vocabulary started to increase its score from 2009, although from 2013 to 2018 it became again steady. At this point in time, the *skos*'s PageRank score is almost half its score in 2003. The *vann* vocabulary had its lowest point in 2010, and started to slowly grow again from 2011. The *grddl* vocabulary appeared in 2008, with the lowest PageRank score, although it was close to *dctype* and *vann*. It slightly decreased in 2009, and then it clearly increased in 2010 to remain almost constant in the following years, with roughly the same value as *dcam*.

Regarding the HITS scores, Figure 5.16a shows that the general trend for all the vocabularies is a rapid increase, albeit with some fluctuations. Specifically, *vann* started to grow from 2007, after an initial slight decrease. In 2018, it achieved the highest authority score. The *skos* vocabulary shows a similar trend, with a more pronounced initial decrease from 2003 to 2004, and a peak in 2011. Subsequently, there was almost no further growth. Notably, *vs*'s authority and



(a) The in-degree scores.



(b) The out-degree scores.

Figure 5.13: The in- and out-degree scores for the top-five vocabularies on each NeLO snapshot after excluding the meta-vocabularies.

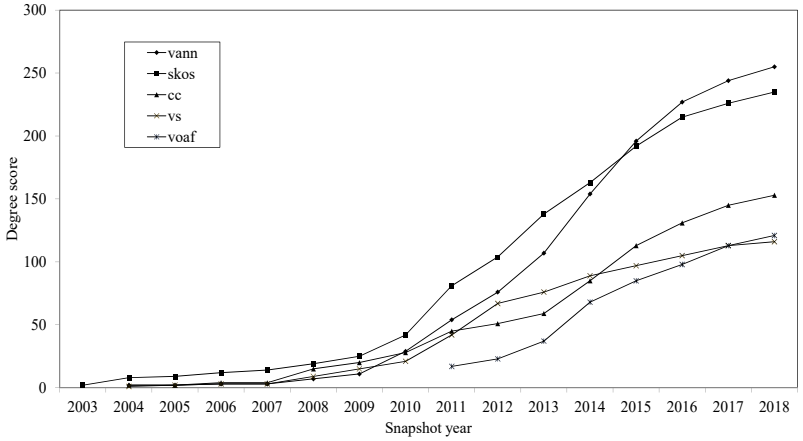


Figure 5.14: The degree scores for the top-five vocabularies on each NeLO snapshot after excluding the meta-vocabularies.

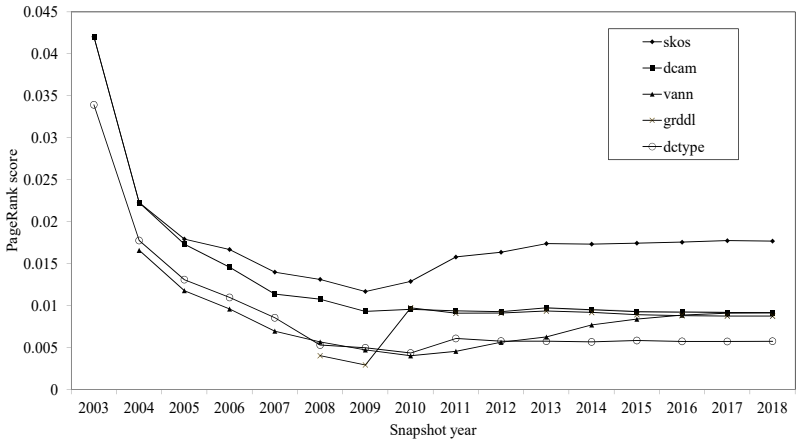


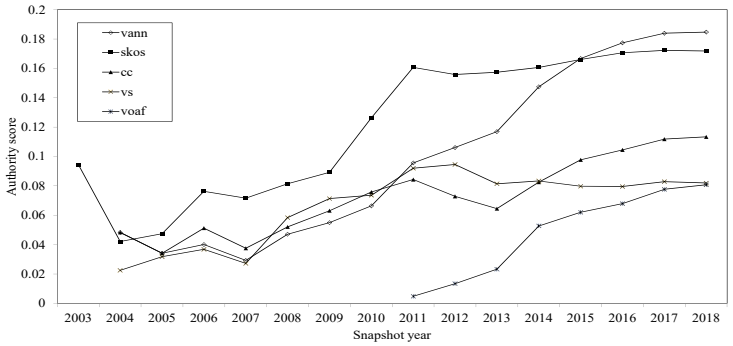
Figure 5.15: The PageRank scores for the top-five vocabularies on each NeLO snapshot after excluding the meta-vocabularies.

hub scores decrease starting from 2013, and then the scores become stable. The *voaf* vocabulary appeared in 2011, and steadily grew until 2018, when it achieved the same values as *vs*. The latter has the lowest scores among the vocabularies presented.

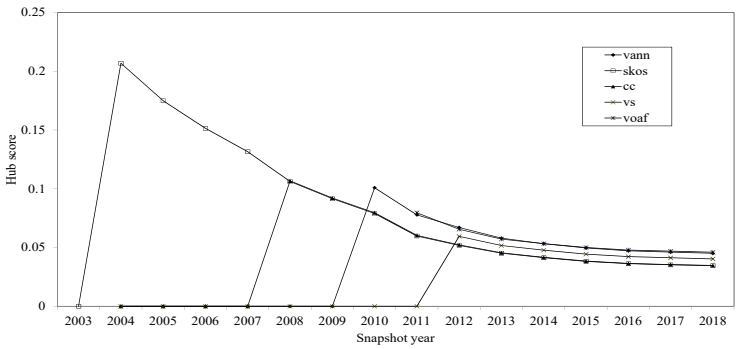
On the other hand, all the previous vocabularies show decreasing hub scores after an initial peak, as shown in Figure 5.16b. The difference is in their peak value and the year. For all vocabularies, the hub score is equal to zero until they quickly reach their peak. In 2018, all the vocabularies achieved similar hub scores of around 0.05. A peak occurs in 2004 for *skos*, in 2008 for *cc*, in 2010 for *vann*, and in 2012 for *vs* and *voaf*.

5.3 Discussion

The statistics of the state of the NeLO in June 2018 shows that there is a need to increase the reuse of types and properties between vocabularies. Due to the fast increase in the number of vocabularies, the number of imports (edges) have decreased over time. Reusing terms from other vocabularies require the ontology engineers to invest more effort to keep track of the changes of the imported types and properties, especially if the number of the imported terms are big. In Section 5.3.1, we discuss the results related to the state of the NeLO, as it appeared in June 2018. The discussion of the evolution of the NeLO over time is presented in Section 5.3.2.



(a) The authority scores.



(b) The hub scores.

Figure 5.16: The HITS scores for the top-five vocabularies on each NeLO snapshot after excluding the meta-vocabularies.

5.3.1 Network of Linked Vocabularies in 2018

In June 2018, NeLO consisted of 994 vocabularies and 7,046 edges between those vocabularies, with a density of 0.007 [3]. We can conclude that the actual number of edges in the graph is far from the maximal number of possible edges (corresponding to having an edge to all other nodes for each node in the graph) and the maximal density (which would be equal to one with the maximal number of possible edges). The average degree of the vocabularies in NeLO 2018 is around 7.1, with a standard deviation of about seven.

The vocabularies in NeLO form three categories (the three circles in Figure 5.1, Section 5.2.1). The first one corresponds to vocabularies, including the meta-vocabularies, that export terms to most of the other vocabularies in the network. These vocabularies in the central circle are the most important in NeLO 2018. They are the most popular, in the sense that their terms are highly reused, but updating their terms is critical because of their potential impact on many other vocabularies which reuse their terms. Nevertheless, the changes in vocabularies occur rather rarely. In fact, these meta-vocabularies had, on average, three versions over 17 years. Overall, the vocabularies in this category represent 2% of all vocabularies, export their terms to 71% of other vocabularies, and account for 66% of outgoing links [3]. Vocabularies in the second category still have many edges to other ontologies, but less than the meta-vocabularies. These are also incredibly popular, and updating them could impact various vocabularies. The average number of versions of the second category of vocabularies is around three. Thus, the vocabularies in this category seem to be more stable than the third category. The vocabularies in the middle circle account for around 20% of the outgoing links. These vocabularies represent 13% of the vocabularies, and their terms are reused by 56% of other vocabularies in NeLO 2018. The third category contains rarely reused vocabularies, such as the newcomers or vocabularies that

cover a very specific domain. We believe that the newcomers will remain in the outer circle in the future, since the more general and the meta-vocabularies (the center circle) cover most of the other vocabularies, needs, and we do not expect that there will be many new generic vocabularies in the future.

5.3.2 Changes in the Network of Linked Vocabularies

The number of new vocabularies and relations between them has decreased over time [3]. While in 2003, 55% of the vocabularies were new, this percentage decreased to 4% in 2018. Regarding the edges in 2003, 57% of them were introduced that year. This percentage decreased to 27% in 2009, increased to 43% in 2010, and dropped to 4% in 2018. Ontology engineers keep adding types and properties to their existing vocabularies, rather than introducing new ontologies, in order to fulfill their domain requirements. Therefore, we expect the number of new vocabularies will continue to decrease over time. Given that less new vocabularies have been introduced over time, it is not surprising that lesser import/export links have also been created.

Considering the reuse of terms from 2004 to 2010, the percentage with respect to the available ones ranges between 58% and 22%. This percentage decreased to 10% in June 2017, slightly increasing in 2018, accounting for 16% of the available terms. This suggests that initially reusing terms was more common. One reason could be that since much fewer vocabularies were available, it was easier to be aware of them and reuse their terms. Nevertheless, more specific vocabularies, which are less suitable to be reused, may have been created over time.

Over time, some vocabularies have become more popular, depending on

the growing number of exports to other vocabularies. Excluding the meta-vocabularies, *vann*, *skos*, *cc*, *vs*, and *voaf* are the most popular. By considering the out-degree and centrality measures on all NeLO snapshots for the vocabularies with the highest scores in the last three years, we found that *vann*, *skos*, *cc*, *vs*, and *voaf* increased their scores. Notably, *vann* and *skos* saw a more rapid increase in scores than the other three vocabularies, i. e., they became popular faster over time. Overall, the meta-vocabularies, which are suitable for most domains, are the most popular ones. Interestingly, our findings show a decline in the growth of out-degree scores, i. e., the average number of exports per vocabulary. This could be due to the fact that less new vocabularies have been introduced over time, consequently, fewer terms are exported to those new vocabularies. Nevertheless, the reuse of terms could still be increased among existing vocabularies, according to the needs of a particular application scenario.

Regarding the in-degree scores, we observed that they vary among the nodes in the network over time. Some of the vocabularies with the highest in-degree over time, such as *mo*, *interval*, and *semio*, show sudden and large growth in imports, at a specific point in time. This corresponds to a new version with a considerable extension of the previous vocabulary, which is then reusing more terms from other ontologies. Thus, subsequently, more effort is needed to keep track of the changes in the imported terms.

Similarly, the changes in the vocabularies with high PageRank and HITS scores affect many other vocabularies. The difference to those with a high out-degree is that their changes can also significantly impact ontologies not directly linked, i. e., their effect is potentially less local. PageRank and HITS can help to identify nodes which can be problematic due to several dependencies. A node with high PageRank and HITS scores mean that this node has many direct and

indirect importers. Thus, a change in this node affects many other nodes in the network. Therefore, these changes can be even more critical, and we recommend that ontologies engineers of the vocabularies that import terms from those with high PageRank and HITS scores periodically check for changes in the imported ones.

5.4 Summary

In this chapter, we analyzed the evolution of the NeLO using the available versions of the vocabularies from LOV, from over more than 17 years (RQ3 Section 1.2). We presented static parameters of NeLO, such as its size, density, average degree, and the most important vocabularies at a certain point of time. We further investigated how the NeLO changes over time. In this regard, we measured the impact of a change in one vocabulary on the other vocabularies in the NeLO. Our analyses provide, for the first time, in-depth insights into the structure and evolution of the NeLO.

From the figure of NeLO as of June 2018, we can conclude that the vocabularies are organized into three categories. A small inner circle that mostly consists of the meta-vocabularies. The middle circle includes the vocabularies that are very popular, but not like the meta-vocabularies. The outer circle contains the rarely used vocabularies and the newcomers. From the NeLO's 2018 density and average degree, we think that there is a need to increase the imports/exports relations between vocabularies to avoid redundancy in terms. Due to a large number of vocabularies in the NeLO, the imports between them have decreased over time, which is expected but not advised. Based on the out-degree scores for the vocabularies in NeLO, some vocabularies have become

more popular over time. This means that they moved from the outer circle to the middle one, and some of those vocabularies come close to the inner circle of the meta-vocabularies, such as *vann*, *skos*, and *cc*. The changes in one vocabulary affect vocabularies that import terms from it. The vocabularies with high PageRank and HITS scores affect the vocabularies that reuse types or properties from them, i. e., they are more critical, and the ontology engineers need to put more effort in to keep track of the changes for reusing terms. Recommender systems for data modeling such as TermPicker [66] can play a major role in helping ontology engineers to select types and properties from other vocabularies.

Chapter 6

Conclusion and Outlook

This thesis presents a comprehensive analysis of the vocabularies of the LOD cloud. We studied the changes of vocabularies and the reuse of vocabulary types and properties by other vocabularies. Furthermore, we analyzed the use of types and properties by data publishers and how the changed terms are adopted in data. Additionally, we studied the NeLO, and how it has evolved over time. Finally, we analyzed how the changes in NeLO's vocabularies can affects the vocabularies that have an import relation to the changed terms.

Below, we draw the main conclusions for each of the three analyses that we conducted to study the evolution of vocabularies on the Semantic Web (Section 6.1). The first analysis studied the vocabulary changes and reuse among vocabularies. The second analysis studied the use of types and properties in data and the adoption of the changed terms by data publishers. The third analysis studied the NeLO and its evolution. In Section 6.2, we present the lessons learned while conducting the analysis of this thesis. The outlook and future directions are illustrated in Section 6.3.

6.1 Conclusions from the Analyses

Vocabulary Changes and Reuse

This work provides a comprehensive analysis of vocabulary dependencies (imported vocabularies) as well as the relation between the changed vocabularies and the vocabularies that import their terms. Changes are mostly made on the terms owned by vocabularies, compared to the changes of the imported terms. Furthermore, most of the vocabularies almost have a fixed number of external vocabularies importing terms from them during the evolution period. If they add or remove some vocabularies, the number of those additions and deletions is small. Vocabulary's domains such as publications, geographic, social web, and government have a high percentage of change compared to other domains such as life sciences and media. Providing support tools to regularly check for updates of vocabularies helps update the vocabularies regularly, rather than checking for updates manually, which may lead to misses.

There is a need to increase the amount of reused types and properties between vocabularies. Based on the vocabularies listed in LOV, only 16% of the existing terms are reused by other vocabularies. Recommender systems can increase this amount by giving information regarding the current terms. Since some of the deprecated and deleted terms are still reused, we think that the process for checking for updates is usually done manually. Furthermore, there is a need for tools that help ontology engineers to keep track of the changes in vocabularies.

Use and Adoption of Vocabulary Terms for Modeling Data

Even small changes in vocabulary terms have an impact on the published data that use those terms. It is of positive significance to observe that most of the newly coined terms are adopted immediately. We found that 50 %, and 23 % of the types and properties studied are never used in the data crawled in the BTC and DyLDO datasets, respectively. Unexpectedly, some deprecated terms were recreated after they became deprecated in prior versions. We are not surprised that most of the deprecated terms are still used because data publishers may not be aware of the changes to the exploited vocabularies. We think that this work can help data publishers in updating their data by raising awareness that some terms that they use have been deleted or deprecated. Providing a service to notify about changes on ontologies can simplify the update of vocabularies and datasets, as well as foster the adoption of new terms. We also believe that raising awareness in data publishers about the existence of new terms in an ontology may further stimulate the use of terms. There exist some vocabularies, such as *vs*, that provide information about the status of the types and properties of the vocabularies. We propose that ontology engineers use them in order to explicitly provide information about the current status of the terms in their vocabularies. Therefore, when data publishers use those types and properties, they know the status of the terms.

Analysis of the Network of Linked Vocabularies

Based on the analysis of the NeLO as of June 2018, we can conclude that the vocabularies are organized into three categories. The inner circle that

mostly consists of the meta-vocabularies. The middle circle includes the popular vocabularies, but not as much as the meta-vocabularies. The outer circle contains the rarely used vocabularies and newcomers. We believe that future newcomers will remain in the outer circle, since the more general vocabularies and meta-vocabularies (the center circle) cover most of the other vocabulary needs, and we do not expect that there will be many new generic vocabularies in the future. As our analysis of the evolution of the NeLO shows, the dynamics of changes slowed down after some fast evolution between 2001 and 2010. This is expected because the domains are almost covered with vocabularies. But we need to increase the reuse to avoid redundancy of terms as much as possible. The vocabularies with high PageRank and HITS scores affect vocabularies that import terms from them, i. e., they are more critical, and the ontology engineers need to put in more effort to keep track of the changes for the terms they reuse from those vocabularies. Thus, with this work, we aim to raise ontology engineers' awareness about the changes in the NeLO and stimulate a further increase of reusing of terms. Analyzing the changes of vocabularies can help ontology engineers in establishing new ontologies.

6.2 Lessons Learned

There are some challenges that we faced during this work. First, finding suitable dataset(s) to conduct the experiments. Second, the availability of vocabularies and their prior versions. In the following paragraphs, a detailed description of these challenges is provided and how we dealt with them.

Findning Datasets. We faced a challenge in finding a suitable dataset to conduct the experiments, especially for the study of analyzing the use of

vocabulary types and properties, and the adoption of vocabulary changes in data. To analyze the use and adoption of vocabulary terms over time, we needed snapshots of data from previous years. The best case for us was if we found snapshots of data from the early years from when data publishers start modeling their data using vocabularies. The early versions of vocabularies were published around 19 years ago. Unfortunately, we could not find a dataset with this specification; thus, we used the DyLDO and BTC datasets. DyLDO started crawling since April 2012, and BTC had yearly snapshots since 2009. To address this challenge, we conducted our experiments on each dataset separately. For DyLDO, we used data starting from the first snapshot. For BTC, we used five data sets, from 2009 to 2012 as well as 2014.

Availability of Versions. While studying the evolution of vocabularies, we needed all or most of the prior versions of the vocabularies. We used the LOV dataset to download and investigate the evolution of vocabularies. The LOV provides a history for the prior versions of vocabularies. The challenge was that we did not find all the prior versions for all vocabularies. We wanted to conduct the analysis by including all possible versions of the vocabularies. The LOV contains only the latest versions for *owl*, *rdf*, and *rdfs*. To address this problem, we searched for other sources of those missing prior versions, i. e., the official portals of vocabularies. For example, we extracted the changes between versions for those three vocabularies, using the documentation in the official websites of those vocabularies. This process took a lot of time and effort to record the changes in vocabularies.

6.3 Outlook

The NeLO is dynamic and continues to grow. This continuous change can motivate to further investigate evolving vocabularies and their relation to linked vocabularies, and the published data. So far, no feedback system tells ontology engineers systematically how their terms are used for describing real data. Likewise, data publishers are not notified when new vocabulary versions appear. To address the challenge of better understanding and also stimulating the adoption of added terms and discontinuing the use of deprecated terms, establishing an online vocabulary system to track the history of types and properties is required.

Furthermore, there is a need to study the impact of vocabulary changes on the ontology network. When a vocabulary changes, it has an impact on the vocabularies that import terms from the changed vocabulary. Changes such as deletion is critical to the vocabularies that import it, since those importers use outdated terms. Another research direction can involve considering the other types of updates, such as adding/removing constraints to terms, subclasses, subproperties, as well as domain and range information. In this work, we focused on two types of changes, creating and deleting/deprecating terms. Vocabularies can be classified according to the domains they represent. Consequently, analyzing the vocabulary terms based on the different domains is another interesting direction of research.

Bibliography

[1] Mohammad Abdel-Qader and Ansgar Scherp. Qualitative analysis of vocabulary evolution on the linked open data cloud. In Elena Demidova, Stefan Dietze, Julian Szymanski, and John G. Breslin, editors, *Proceedings of the 3rd International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES '16) co-located with the 13th ESWC 2016 Conference, Anissaras, Greece, May 30, 2016*, volume 1597 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[2] Mohammad Abdel-Qader, Ansgar Scherp, and Iacopo Vagliano. Analyzing the evolution of vocabulary terms and their impact on the LOD cloud. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 1–16. Springer, 2018.

[3] Mohammad Abdel-Qader, Iacopo Vagliano, and Ansgar Scherp. Analyzing the evolution of linked vocabularies. In Maxim Bakaev, Flavius Frasincar, and In-Young Ko, editors, *Web Engineering - 19th International Conference, ICWE 2019, Daejeon, South Korea, June 11-14, 2019,*

Proceedings, volume 11496 of *Lecture Notes in Computer Science*, pages 409–424. Springer, 2019.

[4] Charu C. Aggarwal and Nan Li. On node classification in dynamic content-based networks. In *Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM 2011, April 28-30, 2011, Mesa, Arizona, USA*, pages 355–366. SIAM / Omnipress, 2011.

[5] Usman Akhtar, Muhammad Asif Razzaq, Ubaid Ur Rehman, Muhammad Bilal Amin, Wajahat Ali Khan, Eui-Nam Huh, and Sungyoung Lee. Change-aware scheduling for effectively updating linked open data caches. *IEEE Access*, 6:65862–65873, 2018.

[6] Jørgen Bang-Jensen and Gregory Z. Gutin. Basic terminology, notation and results. In Jørgen Bang-Jensen and Gregory Z. Gutin, editors, *Classes of Directed Graphs*, Springer Monographs in Mathematics, pages 1–34. Springer, 2018.

[7] Vladimir Batagelj and Andrej Mrvar. Pajek - analysis and visualization of large networks. In Petra Mutzel, Michael Jünger, and Sebastian Leipert, editors, *Graph Drawing, 9th International Symposium, GD 2001 Vienna, Austria, September 23-26, 2001, Revised Papers*, volume 2265 of *Lecture Notes in Computer Science*, pages 477–478. Springer, 2001.

[8] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The Semantic Web. *Scientific American*, 284(5):28–37, 2001.

[9] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[10] Ulrik Brandes and Steven R. Corman. Visual unrolling of network evolution and the analysis of dynamic discourse? *Information Visualization*, 2(1):40–50, 2003.

[11] Dan Brickley, Ramanathan V Guha, and Brian McBride. RDF vocabulary description language 1.0: RDF schema. w3c recommendation (2004). *URL http://www. w3. org/tr/2004/rec-RDF-schema-20040210*, 2004.

[12] Silvio Domingos Cardoso, Cédric Pruski, Marcos Da Silveira, Ying-Chi Lin, Anika Groß, Erhard Rahm, and Chantal Reynaud-Delaître. Leveraging the impact of ontology evolution on semantic annotations. In Eva Blomqvist, Paolo Ciancarini, Francesco Poggi, and Fabio Vitali, editors, *Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings*, volume 10024 of *Lecture Notes in Computer Science*, pages 68–82, 2016.

[13] Rathachai Chawuthai, Hideaki Takeda, Vilas Wuwongse, and Utsugi Jinbo. Presenting and preserving the change in taxonomic knowledge for linked data. *Semantic Web*, 7(6):589–616, 2016.

[14] Thomas F. Coleman and Jorge J. Moré. Estimation of sparse hessian matrices and graph coloring problems. *Math. Program.*, 28(3):243–270, 1984.

[15] Gabor Csardi, Tamas Nepusz, et al. The igraph software package for complex network research. *InterJournal, complex systems*, 1695(5):1–9, 2006.

[16] Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised prediction of citation influences. In Zoubin Ghahramani, editor, *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, pages 233–240. ACM, 2007.

[17] Renata Queiroz Dividino, Thomas Gottron, and Ansgar Scherp. Strategies for efficiently keeping local linked open data caches up-to-date. In Marcelo Arenas, Óscar Corcho, Elena Simperl, Markus Strohmaier, Mathieu d'Aquin, Kavitha Srinivas, Paul Groth, Michel Dumontier, Jeff Heflin, Krishnaprasad Thirunarayan, and Steffen Staab, editors, *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II*, volume 9367 of *Lecture Notes in Computer Science*, pages 356–373. Springer, 2015.

[18] Renata Queiroz Dividino, Thomas Gottron, Ansgar Scherp, and Gerd Gröner. From changes to dynamics: Dynamics analysis of linked open data sources. In Elena Demidova, Stefan Dietze, Julian Szymanski, and John G. Breslin, editors, *Proceedings of the 1st International Workshop on Dataset PROFIling & fEderated Search for Linked Data co-located with the 11th Extended Semantic Web Conference, PROFILES@ESWC 2014, Anissaras, Crete, Greece, May 26, 2014*, volume 1151 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[19] Renata Queiroz Dividino, Ansgar Scherp, Gerd Gröner, and Thomas Grotton. Change-a-lod: Does the schema on the linked data cloud change or not? In Olaf Hartig, Juan F. Sequeda, Aidan Hogan, and Takahide Matsutsuka, editors, *Proceedings of the Fourth International Workshop on Consuming Linked Data, COLD 2013, Sydney, Australia, October 22, 2013*,

volume 1034 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[20] Uriel Feige. On sums of independent random variables with unbounded variance and estimating the average degree in a graph. *SIAM J. Comput.*, 35(4):964–984, 2006.

[21] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[22] Amir Ghazvinian, Natalya Fridman Noy, Clément Jonquet, Nigam H. Shah, and Mark A. Musen. What four million mappings can tell you about two hundred ontologies. In Abraham Bernstein, David R. Karger, Tom Heath, Lee Feigenbaum, Diana Maynard, Enrico Motta, and Krishnaprasad Thirunarayan, editors, *The Semantic Web - ISWC 2009, 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings*, volume 5823 of *Lecture Notes in Computer Science*, pages 229–242. Springer, 2009.

[23] Thomas Gottron and Christian Gottron. Perplexity of index models over evolving linked data. In Valentina Presutti, Claudia d'Amato, Fabien Gandon, Mathieu d'Aquin, Steffen Staab, and Anna Tordai, editors, *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, volume 8465 of *Lecture Notes in Computer Science*, pages 161–175. Springer, 2014.

[24] Thomas Gottron, Malte Knauf, and Ansgar Scherp. Analysis of schema structures in the linked open data graph based on unique subject uris,

pay-level domains, and vocabulary usage. *Distributed Parallel Databases*, 33(4):515–553, 2015.

[25] Thomas R Gruber et al. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993.

[26] Ramanathan V. Guha, Dan Brickley, and Steve Macbeth. Schema.org: Evolution of structured data on the web. *ACM Queue*, 13(9):10, 2015.

[27] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using NetworkX. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[28] Andreas Harth, Katja Hose, and Ralf Schenkel, editors. *Linked Data Management*. Chapman and Hall/CRC, 2014.

[29] Michael Hartung, Toralf Kirsten, and Erhard Rahm. Analyzing the evolution of life science ontologies and mappings. In Amos Bairoch, Sarah Cohen Boulakia, and Christine Froidevaux, editors, *Data Integration in the Life Sciences, 5th International Workshop, DILS 2008, Evry, France, June 25-27, 2008. Proceedings*, volume 5109 of *Lecture Notes in Computer Science*, pages 11–27. Springer, 2008.

[30] Pascal Hitzler, Markus Krötzsch, and Sebastian Rudolph. *Foundations of Semantic Web Technologies*. Chapman and Hall/CRC Press, 2010.

[31] Aidan Hogan. Canonical forms for isomorphic and equivalent RDF graphs: Algorithms for leaning and labelling blank nodes. *ACM Trans. Web*, 11(4):22:1–22:62, 2017.

[32] Chuan Hu and Huiping Cao. Discovering time-evolving influence from dynamic heterogeneous graphs. In *2015 IEEE International Conference on Big Data, Big Data 2015, Santa Clara, CA, USA, October 29 - November 1, 2015*, pages 2253–2262. IEEE Computer Society, 2015.

[33] Chuan Hu, Huiping Cao, and Chaomin Ke. Detecting influence relationships from graphs. In Mohammed Javeed Zaki, Zoran Obradovic, Pang-Ning Tan, Arindam Banerjee, Chandrika Kamath, and Srinivasan Parthasarathy, editors, *Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24-26, 2014*, pages 821–829. SIAM, 2014.

[34] Robert Isele, Jürgen Umbrich, Christian Bizer, and Andreas Harth. Ldspider: An open-source crawling framework for the web of linked data. In Axel Polleres and Huajun Chen, editors, *Proceedings of the ISWC 2010 Posters & Demonstrations Track: Collected Abstracts, Shanghai, China, November 9, 2010*, volume 658 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.

[35] Maciej Janik, Ansgar Scherp, and Steffen Staab. The semantic web: Collective intelligence on the web. *Inform. Spektrum*, 34(5):469–483, 2011.

[36] Ernesto Jiménez-Ruiz, Bernardo Cuenca Grau, Ulrike Sattler, Thomas Schneider, and Rafael Berlanga Llavori. Safe and economic re-use of ontologies: A logic-based methodology and tool support. In Sean Bechhofer, Manfred Hauswirth, Jörg Hoffmann, and Manolis Koubarakis, editors, *The Semantic Web: Research and Applications, 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008, Proceedings*, volume 5021 of *Lecture Notes in Computer*

Science, pages 185–199. Springer, 2008.

[37] Tobias Käfer, Ahmed Abdelrahman, Jürgen Umbrich, Patrick O'Byrne, and Aidan Hogan. Observing linked data dynamics. In Philipp Cimiano, Óscar Corcho, Valentina Presutti, Laura Hollink, and Sebastian Rudolph, editors, *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*, volume 7882 of *Lecture Notes in Computer Science*, pages 213–227. Springer, 2013.

[38] Tobias Käfer, Jürgen Umbrich, Aidan Hogan, and Axel Polleres. Dyldo: Towards a dynamic linked data observatory. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Michael Hausenblas, editors, *WWW2012 Workshop on Linked Data on the Web, Lyon, France, 16 April, 2012*, volume 937 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.

[39] Maulik R. Kamdar, Tania Tudorache, and Mark A. Musen. A systematic analysis of term reuse and term overlap across biomedical ontologies. *Semantic Web*, 8(6):853–871, 2017.

[40] Vipul Kashyap, Christoph Bussler, and Matthew Moran. *The Semantic Web - Semantics for Data and Services on the Web*. Data-Centric Systems and Applications. Springer, 2008.

[41] Ravneet Kaur and Sarbjeet Singh. A review of social network centric anomaly detection techniques. *Int. J. Commun. Networks Distributed Syst.*, 17(4):358–386, 2016.

[42] Arijit Khan, Sourav S. Bhowmick, and Francesco Bonchi. Summarizing

static and dynamic big graphs. *Proc. VLDB Endow.*, 10(12):1981–1984, 2017.

[43] Rob Kitchin. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, 2014.

[44] Michel C. A. Klein and Dieter Fensel. Ontology versioning on the semantic web. In Isabel F. Cruz, Stefan Decker, Jérôme Euzenat, and Deborah L. McGuinness, editors, *Proceedings of SWWS'01, The first Semantic Web Working Symposium, Stanford University, California, USA, July 30 - August 1, 2001*, pages 75–91, 2001.

[45] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[46] Graham Klyne. Resource description framework (RDF): Concepts and abstract syntax. *http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/*, 2004.

[47] Hsin-Tsang Lee, Derek Leonard, Xiaoming Wang, and Dmitri Loguinov. Irlbot: Scaling to 6 billion pages and beyond. *ACM Trans. Web*, 3(3):8:1–8:34, 2009.

[48] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In Ying Li, Bing Liu, and Sunita Sarawagi, editors, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 462–470. ACM, 2008.

[49] Jure Leskovec and Rok Sosic. SNAP: A general-purpose network analysis and graph-mining library. *ACM Trans. Intell. Syst. Technol.*, 8(1):1:1–1:20, 2016.

[50] Lu Liu, Jie Tang, Jiawei Han, Meng Jiang, and Shiqiang Yang. Mining topic-level influence in heterogeneous networks. In Jimmy Huang, Nick Koudas, Gareth J. F. Jones, Xindong Wu, Kevyn Collins-Thompson, and Aijun An, editors, *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, pages 199–208. ACM, 2010.

[51] Deborah L McGuinness, Frank Van Harmelen, et al. OWL web ontology language overview. *W3C recommendation*, 10(10):2004, 2004.

[52] Robert Meusel, Christian Bizer, and Heiko Paulheim. A web-scale study of the adoption and evolution of the schema.org vocabulary over time. In Rajendra Akerkar, Marios D. Dikaiakos, Achilleas Achilleos, and Tope Omitola, editors, *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics, WIMS 2015, Larnaca, Cyprus, July 13-15, 2015*, pages 15:1–15:11. ACM, 2015.

[53] Nandana Mihindukulasooriya, María Poveda-Villalón, Raúl García-Castro, and Asunción Gómez-Pérez. Collaborative ontology evolution and data quality - an empirical analysis. In Mauro Dragoni, María Poveda-Villalón, and Ernesto Jiménez-Ruiz, editors, *OWL: - Experiences and Directions - Reasoner Evaluation - 13th International Workshop, OWLED 2016, and 5th International Workshop, ORE 2016, Bologna, Italy, November 20, 2016, Revised Selected Papers*, volume 10161 of *Lecture Notes in Computer Science*, pages 95–114. Springer, 2016.

[54] Joachim Neubert. Leveraging SKOS to trace the overhaul of the STW thesaurus for economics. In Mariana Curado Malta and Silvana Aparecida Borsetti Gregório Vidotti, editors, *Proceedings of the 2015 International Conference on Dublin Core and Metadata Applications, DC 2015, São Paulo, Brazil, September 1-4, 2015*, pages 170–180. Dublin Core Metadata Initiative, 2015.

[55] Chifumi Nishioka and Ansgar Scherp. Temporal patterns and periodicity of entity dynamics in the linked open data cloud. In Ken Barker and José Manuél Gómez-Pérez, editors, *Proceedings of the 8th International Conference on Knowledge Capture, K-CAP 2015, Palisades, NY, USA, October 7-10, 2015*, pages 22:1–22:4. ACM, 2015.

[56] Chifumi Nishioka and Ansgar Scherp. Information-theoretic analysis of entity dynamics on the linked open data cloud. In Elena Demidova, Stefan Dietze, Julian Szymanski, and John G. Breslin, editors, *Proceedings of the 3rd International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES '16) co-located with the 13th ESWC 2016 Conference, Anissaras, Greece, May 30, 2016*, volume 1597 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[57] Mahda Noura, Amelie Gyrard, Sebastian Heil, and Martin Gaedke. Concept extraction from the web of things knowledge bases. In *International Conference WWW/Internet*, 2018.

[58] Natalya Fridman Noy and Mark A. Musen. Ontology versioning in an ontology management framework. *IEEE Intell. Syst.*, 19(4):6–13, 2004.

[59] Naoto Ohsaka, Takuya Akiba, Yuichi Yoshida, and Ken-ichi

Kawarabayashi. Dynamic influence analysis in evolving networks. *Proc. VLDB Endow.*, 9(12):1077–1088, 2016.

[60] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[61] Raúl Palma, Fouad Zablith, Peter Haase, and Óscar Corcho. Ontology evolution. In Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, Enrico Motta, and Aldo Gangemi, editors, *Ontology Engineering in a Networked World*, pages 235–255. Springer, 2012.

[62] Vassilis Papakonstantinou, Irini Fundulaki, and Giorgos Flouris. Assessing linked data versioning systems: The semantic publishing versioning benchmark. In Thorsten Liebig, Achille Fokoue, and Zhe Wu, editors, *Proceedings of the 12th International Workshop on Scalable Semantic Web Knowledge Base Systems co-located with 17th International Semantic Web Conference, SSWS@ISWC 2018, Monterey, California, USA, October 9, 2018*, volume 2179 of *CEUR Workshop Proceedings*, pages 45–60. CEUR-WS.org, 2018.

[63] Yannis Roussakis, Ioannis Chrysakis, Kostas Stefanidis, Giorgos Flouris, and Yannis Stavrakas. A flexible framework for understanding the dynamics of evolving RDF datasets. In Marcelo Arenas, Óscar Corcho, Elena Simperl, Markus Strohmaier, Mathieu d'Aquin, Kavitha Srinivas, Paul Groth, Michel Dumontier, Jeff Heflin, Krishnaprasad Thirunarayan, and Steffen Staab, editors, *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part*

I, volume 9366 of *Lecture Notes in Computer Science*, pages 495–512. Springer, 2015.

[64] Purnamrita Sarkar, Deepayan Chakrabarti, and Michael I. Jordan. Nonparametric link prediction in dynamic networks. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012.

[65] Johann Schaible, Thomas Gottron, and Ansgar Scherp. Survey on common strategies of vocabulary reuse in linked open data modeling. In Valentina Presutti, Claudia d'Amato, Fabien Gandon, Mathieu d'Aquin, Steffen Staab, and Anna Tordai, editors, *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, volume 8465 of *Lecture Notes in Computer Science*, pages 457–472. Springer, 2014.

[66] Johann Schaible, Thomas Gottron, and Ansgar Scherp. Termpicker: Enabling the reuse of vocabulary terms by exploiting data from the linked open data cloud. In Harald Sack, Eva Blomqvist, Mathieu d'Aquin, Chiara Ghidini, Simone Paolo Ponzetto, and Christoph Lange, editors, *The Semantic Web. Latest Advances and New Domains - 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29 - June 2, 2016, Proceedings*, volume 9678 of *Lecture Notes in Computer Science*, pages 101–117. Springer, 2016.

[67] Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the linked data best practices in different topical domains. In Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig A. Knoblock, Denny Vrandecic, Paul Groth, Natasha F. Noy, Krzysztof Janowicz,

and Carole A. Goble, editors, *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, volume 8796 of *Lecture Notes in Computer Science*, pages 245–260. Springer, 2014.

[68] Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the linked data best practices in different topical domains. In Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig A. Knoblock, Denny Vrandecic, Paul Groth, Natasha F. Noy, Krzysztof Janowicz, and Carole A. Goble, editors, *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, volume 8796 of *Lecture Notes in Computer Science*, pages 245–260. Springer, 2014.

[69] Thanos G. Stavropoulos, Stelios Andreadis, Efstratios Kontopoulos, and Ioannis Kompatsiaris. Semadrift: A hybrid method and visual tools to measure semantic drift in ontologies. *J. Web Semant.*, 54:87–106, 2019.

[70] ChunYuen Teng, Liuling Gong, Avishay Livne, Celso Brunetti, and Lada A. Adamic. Coevolution of network structure and content. In Noshir S. Contractor, Brian Uzzi, Michael W. Macy, and Wolfgang Nejdl, editors, *Web Science 2012, WebSci '12, Evanston, IL, USA - June 22 - 24, 2012*, pages 288–297. ACM, 2012.

[71] Jürgen Umbrich, Marcel Karnstedt, and Sebastian Land. Towards understanding the changing web: Mining the dynamics of Linked-Data sources and entities. In Martin Atzmueller, Dominik Benz, Andreas Hotho, and Gerd Stumme, editors, *LWA 2010 - Lernen, Wissen & Adaptivität, Workshop Proceedings, Kassel, 4.-6. Oktober 2010*, pages 159–162, 2010.

[72] Pierre-Yves Vandenbussche, Ghislain Atemezing, María Poveda-Villalón, and Bernard Vatant. Linked open vocabularies (LOV): A gateway to reusable semantic vocabularies on the web. *Semantic Web*, 8(3):437–452, 2017.

[73] Simon Walk, Philipp Singer, Lisette Espin Noboa, Tania Tudorache, Mark A. Musen, and Markus Strohmaier. Understanding how users edit ontologies: Comparing hypotheses about four real-world projects. In Marcelo Arenas, Óscar Corcho, Elena Simperl, Markus Strohmaier, Mathieu d'Aquin, Kavitha Srinivas, Paul Groth, Michel Dumontier, Jeff Heflin, Krishnaprasad Thirunarayan, and Steffen Staab, editors, *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I*, volume 9366 of *Lecture Notes in Computer Science*, pages 551–568. Springer, 2015.

[74] Simon Walk, Philipp Singer, Markus Strohmaier, Denis Helic, Natalya Fridman Noy, and Mark A. Musen. How to apply markov chains for modeling sequential edit patterns in collaborative ontology-engineering projects. *Int. J. Hum. Comput. Stud.*, 84:51–66, 2015.

[75] Fouad Zablith, Grigoris Antoniou, Mathieu d'Aquin, Giorgos Flouris, Haridimos Kondylakis, Enrico Motta, Dimitris Plexousakis, and Marta Sabou. Ontology evolution: a process-centric survey. *Knowl. Eng. Rev.*, 30(1):45–75, 2015.

[76] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93, 2016.